Assignment 5 — Due Nov 27, 2018

NAIQING CAI

ncai5@wisc.edu

1. Eight isolates of rose blackspot fungus

(a) Consider an additive model Y ∼ Isolate + Temp. Complete the decomposition of the total sum of squares:

By running r function, we can get the table as followed:

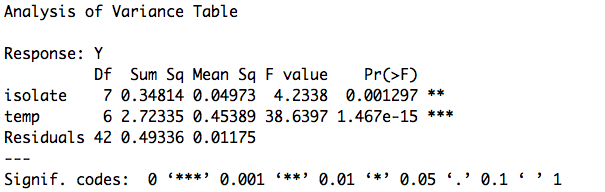


Figure 1.1 Analysis of Variance Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | SS | df | MS | F | p-value |
| Isolate | 0.35 | 7 | 0.0497 | 4.2338 | 0.001 |
| Temp | 2.72 | 6 | 0.4539 | 38.6397 | 1.467e-15 |
| Residual | 0.49 | 42 | 0.01175 |  |  |

Table 1.2 Anova table for model Y ∼ Isolate + Temp

(b) Now consider a linear model where the covariates are polynomials of temperature.

First, fit the linear regression model:

lm0=lm(Y~1)

lm1=lm(Y~1+x)

lm2=lm(Y~1+x+x2)

lm3=lm(Y~x+x2+x3)

lm4=lm(Y~x+x2+x3+x4)

lm6=lm(Y~x+x2+x3+x4+x5+x6)

Then by using anova on either two linear regression models, we can get the table.

(P1|1)=anova(lm0,lm1)

(P2|P1)=anova(lm1,lm2)

(P3|P2)=anova(lm2,lm3)

(P4|P3)=anova(lm3,lm4)

(TEMP|P4)=anova(lm4,lm6)

(TEMP|P3)=anova(lm3,lm6)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| source | SS | DF | MS | F | P-VALUE |
| P1|1 | 0.0117 | 1 | 0.0117 | 0.1781 | 0.6747 |
| P2|P1 | 2.56781 | 1 | 2.56781 | 138.12 | < 2.2e-16 |
| P3|P2 | 0.09992 | 1 | 0.09992 | 5.8683 | 0.01894 |
| P4|P3 | 0.01345 | 1 | 0.01345 | 0.7864 | 0.3793 |
| TEMP(P6)|P4 | 0.03045 | 2 | 0.01523 | 0.8867 | 0.4185 |
| TEMP|P3 | 0.0439 | 3 | 0.01463 | 0.8521 | 0.4722 |

Table 1.3 Anova table for linear model

(c) What does the preceding table tell you about the effect of temperature on growth? Estimate the temperature at which the growth rate is a maximum.

Based the preceding table, we can find that the model: Y~x+x^2+x^3 is the best regression model since its p-value is small.

By optimize() function, we can calculate that temperature =71.74756, the growth rate is a maximum reached 1.05296

2. Flushot

(a) Fit a multiple logistic regression model with the three explanatory variables by maximum likelihood estimation. Report the summary table from the R output. Obtain and interpret the maximum likelihood estimates of β0, β1, β2, and β3. State the fitted logistic response function.

**Multiple Logistic Regression Model**



X1: age

X2: health awareness index, for which higher values indicate greater awareness

X3: gender, where males were coded X3 = 1 and females were coded X3 = 0

**R summary table**

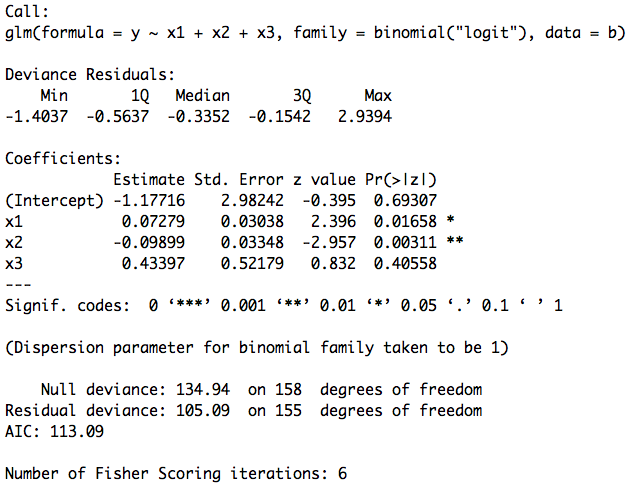


Table 2.1 R summary table

**Interpretation**

β0=-1.17716, which means the average odds of a person having a flu shot is -1.17716..

β1=0.07279, representing the odds of a person having a flu shot increase by about 7.3% with each unit increase in age.

β2=-0.09899, representing the odds of a person having a flu shot decrease by about 9.9% with each unit increase in awareness index.

β3=0.43397, representing the difference between the log odds for male and the log odds for female is 0.43397.

**Fitted Logistic Response Function**

****

(b) Obtain exp(β1), exp(β2), and exp(β3) and the respective 95% CI. Interpret the results.

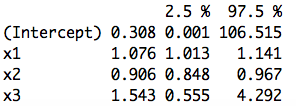


Table 2.2 Result of Confidence Interval

**Interpretation**

exp(β1)=1.076, representing estimated odds ratio for X1

95% CI for exp(β1) is (1.013,1.141)

exp(β2)=0.906, representing estimated odds ratio for X2

95% CI for exp(β2) is (0.848,0.967)

exp(β3)=1.543, representing estimated odds ratio for X3

95% CI for exp(β3) is (0.555,4.292)

(c) What is the estimated probability that male clients aged 55 with a health awareness index of 60 will receive a flu shot? Provide a 95% CI. Interpret the results.

**Estimated Probability**





The result shows that male clients aged 55 with a health awareness index of 60 receive a flu shot at the probability 6.4%.

**95% CI**







The result shows that male clients aged 55 with a health awareness index of 60 have 95% probability to receive a flu shot at the probability between 0.0247 and 0.1568.

(d) Perform a **Wald test** to determine whether X3, client gender, can be dropped from the regression model.

**Multiple Logistic Regression Model:**



**Hypothesis:**



**Wald test Statistic:**





**Conclusion:**

We accept null hypothesis, that is β3=0 and thus X3 can be dropped from the regression model.

(e) Perform a **likelihood ratio test** to determine whether X3, client gender, can be dropped from the regression model.

**Multiple Logistic Regression Model:**

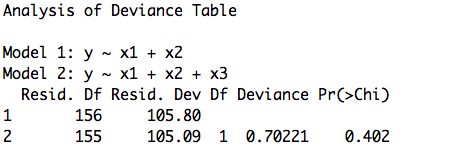


**Hypothesis:**



**Likelihood ratio test Statistic:**



  
Table 2.3 Likelihood Radio Test

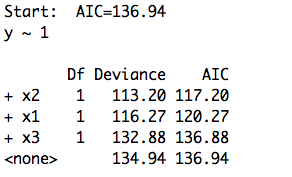


Thus, based on the rejection rule, we should reject null hypothesis and beta3=0 could be dropped from the model.

(f) Perform stepwise model selection based on AIC. Report the summary table for your final model.

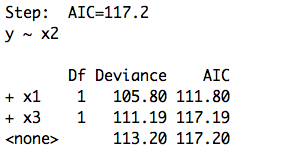
(1) Forward stepwise+ AIC

**Step 1**



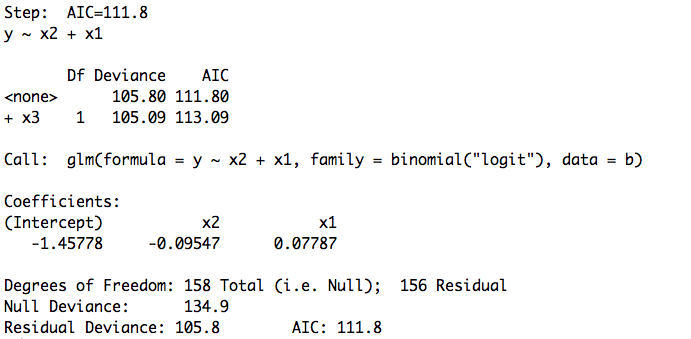
Model start with y~1. Based on the step one, AIC=136.94. These three variables AIC are all greater than 136.94, but we should first add x2 into the model since its AIC is the smallest and thus it has most effect on the model. Then the model becomes y~x2.

**Step 2**



After add x2 as a variable into the model, based on the step two, AIC=117.2. We should then add x1 into the model since its AIC is smaller than 117.2 and it is the smallest and thus it has most effect on the model. Then the model becomes y~x2+x1.

**Step 3**



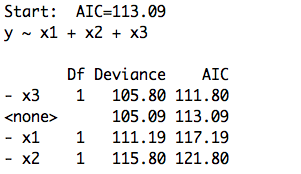
Now, the model becomes y~x2+x1, and then after step three, AIC=111.8 and we find that x3’s AIC is greater than 111.8, so it does not have much effect on the model. We will not add it into the model.

**Finally, we get the model**



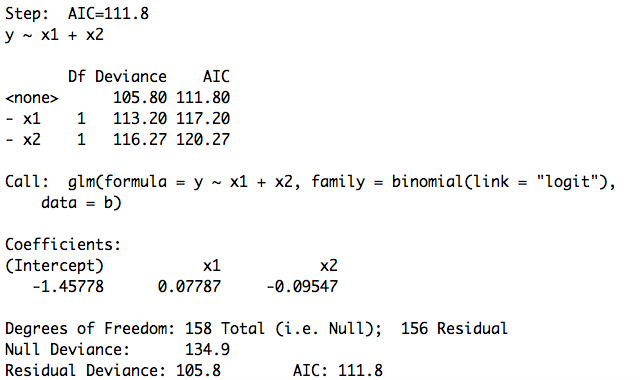
(2) Backward stepwise+ AIC

**Step 1**



Model start with y~x1+x2+x3. Based on the step one, AIC=113.09. x3’s AIC is 111.8 which is smaller than 113.09, so we should first remove x3 from the model since its AIC is the smallest and thus it has least effect on the model. Then the model becomes y~x1+x2.

**Step 2**



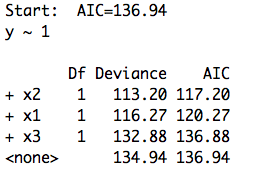
Now, the model is y~x1+x2, and AIC=111.8. x1 and x2’s AIC are both greater than 111.8, as they both don’t need to be removed from the model. And the model is still y~x1+x2.

**Finally, we get the model**



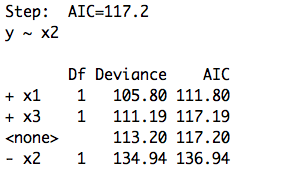
(3) Both direction selection+ AIC

**Step 1**



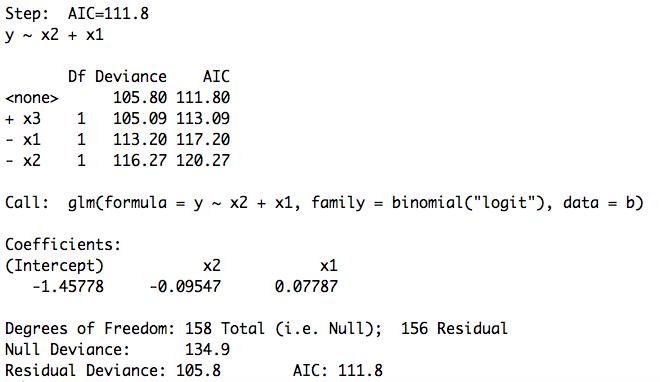
Based on the step one, AIC=136.94. These three variables AIC are all greater than 136.94, but we should first add x2 into the model since its AIC is the smallest and thus it has most effect on the model. Then the model becomes y~x2.

**Step 2**



After add x2 as a variable into the model, based on the step two, AIC=117.2. We should then add x1 into the model since its AIC is smaller than 117.2 and it is the smallest and thus it has most effect on the model. Then the model becomes y~x2+x1.

**Step 3**



Now, the model becomes y~x2+x1, and then after step three, AIC=111.8 and we find that x3’s AIC is greater than 111.8, so it does not have much effect on the model. We will not add it into the model.

**Finally, we get the model**

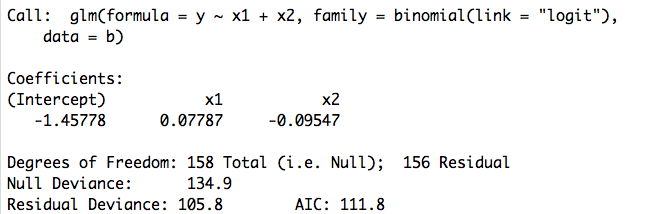
****

Table 2.4 R summary table for final model



(g) Create an ROC plot for the final model. Interpret the results.

hw5%20figure/roc.pdf

Figure 2.5 ROC Plot

Area under the curve: 0.8094

The result shows excellent discrimination, which means that the probability that predictions and the outcomes are concordant is 80.94%.

3. In a clinical trial, m subjects are assigned to each of the treatment group and the control group. In the treatment group, y1 of the m subjects have positive response, while in the control group, y2 subjects have positive response. We are interested in estimating the treatment effect and providing its confidence interval.

(a) Present the above problem as a logistic regression problem and give the interpretation of the regression coefficients β.

|  |  |  |
| --- | --- | --- |
| Y | control group(Xi=0) | treatment group(Xi=1) |
| # of positive response | y2 | y1 |
| # of negative response | m-y2 | m-y1 |
| total | m | m |









**Logistic Regression Model**



**Interpretation of the regression coefficients β**



β0: representing the average odds of of response variable corresponding to the random samples from control group.

β1: representing the difference between the log odds for subjects from treatment group and the log odds for subjects from control group.

(b) Express the likelihood function as a function of β, and derive the MLE βˆ.

**Likelihood Function**









**MLE βˆ**







(c) Find the asymptotic variance-covariance matrix of βˆ.











(d) Suppose that m = 20, y1 = 12, and y2 = 9. Give the point estimate and the corresponding 95% confidence interval for the odds ratio of having positive response between the treatment group and the control group. Can you conclude that the treatment is effective?













From the result we can conclude that the treatment is not effective.

4. Flour beetles Tribolium castaneum were sprayed with one of three insecticides in solution at different doses. The number of insects killed after a six-day period is recorded below

(a) Investigate graphically the relationship between the dose, either in original units or in log units, and the kill rate.

logdose~killrate.pdf

Figure 4.1 Relationship between the doses (log units) and the kill rate

Based on the graph above, the dose and kill rate are positively related.

(b) On the graph for part (a), plot the linear logistic fitted curve for each of the insecticides plus the combination.

I fit the logistic regression and I will have the model:

Rplot01.pdf

Figure 4.2 Linear Logistic Fitted Curve

(c) Consider the two models, one in which the relationship is described by three parallel straight lines in the log dose and and one in which the three lines are straight but not parallel. Assess the evidence against the hypothesis of parallelism.

**Model one:**









**Model two:**



**Assess the evidence against the hypothesis of parallelism：**

 and 

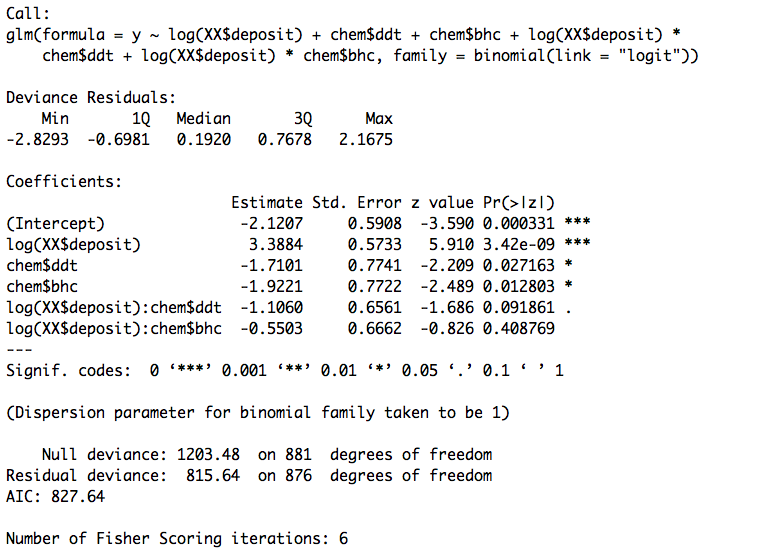


Figure 4.3 Anova summary table for parallelism

Based on the result, we find that p-value is 0.09186 and 0.408769 respectively, they are both >0.025. So we can conclude that we should accept the null hypothesis that beta4=0 and beta5=0 which means three straight lines are parallel.

(d) Let chem denote a 3-level categorical factor, and let ldose be the log dose. Explain the relationship between the regression coefficients in the model formulae chem + ldose and chem + ldose - 1. Explain the relationship between the two covariance matrices.

**Model 1:** lm.reg1=lm(killrate~chem+ldose)

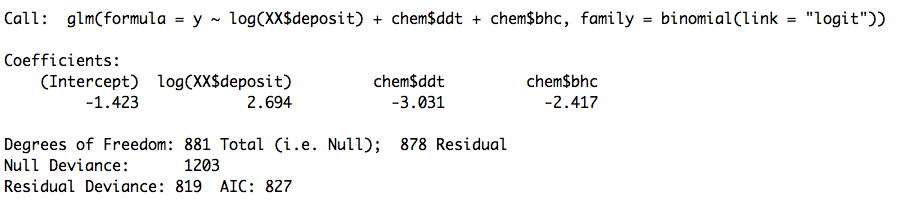


Figure 4.4 model 1 regression

Based on the table, we can find that

regression coefficient for intercept is -1.423

regression coefficient for chemDDT is -3.031

regression coefficient for chemBHC is -2.417

regression coefficient for deposit is 2.694

**Model 2:** lm.reg2=lm(killrate~chem+ldose-1)

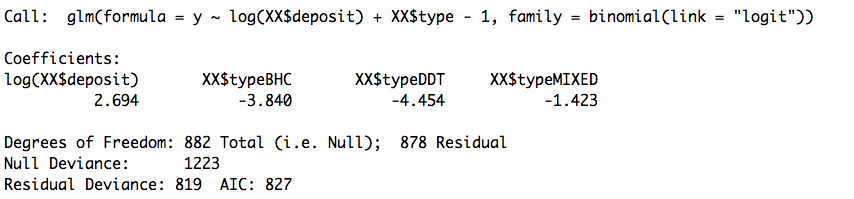


Figure 4.5 model 2 regression

Based on the table, we can find that

regression coefficient for chemDDT is -4.454

regression coefficient for chemBHC is -3.840

regression coefficient for chemMIXED is -1.423

regression coefficient for deposit is 2.694

**Comparison**

****

****

Regression coefficient for deposit is the same for two models : bata1=alpha0

For DDT: beta0+bata2=alpha2

For BHC: bata0+bata3=alpha1

For MIXED: beta0=alpha3

So, in this sense -1 remove the intercept, change it into MIXED term, and put the intercept into terms DDT and BHC.

**Covariance Matrix**

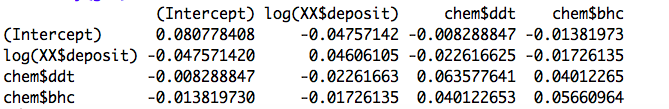


Figure 4.6 Covariance Matrix for model 1

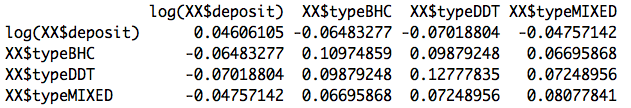


Figure 4.7 Covariance Matrix for model 2

**Relationship**

Actually, the relationship between the two covariance matrices is a kind of linear transformation.

****

****





(e) On the assumption that three parallel straight lines suffice, estimate the potency of the combi- nation relative to each of the components. Obtain a 90% confidence interval for each of these relative potencies.

**Model**









**Combination**



**Components(ddt&bhc)**

****

****

**Potency**







So, we need to estimate the coefficient beta1 and beta2.

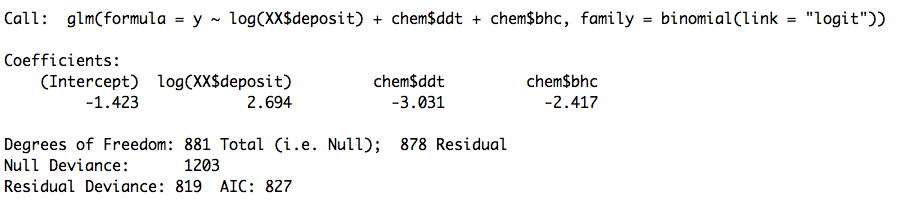


Figure 4.8 R summary table for general linear logistic regression

By r function, beta1=-3.0314, beta2=-2.4170.

Potency of the combination relative to DDT is exp(3.0314)= 20.72623.

Potency of the combination relative to BHC is exp(2.4170)= 11.21217.

**90% CI**

90% CI for beta1 (-3.455892, -2.6257107)

90% CI for beta2 (-2.816464, -2.0330715)

90% CI for potency of the combination relative to DDT is (13.81439, 31.68654)

90% CI for potency of the combination relative to BHC is (7.637509, 16.71763)

(f) Check to see if one of the alternative link functions probit, c-log log or log log, gives an appreciably better fit. Give the answer to part (e) for the c-log log model.

**1) link =probit**

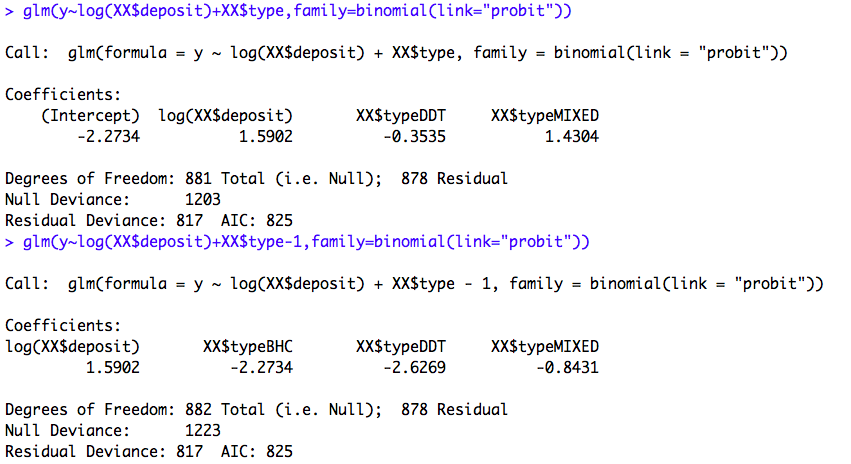


Figure 4.9 R summary table for link=probit

When link =probit, AIC=825 which is smaller than link=logit AIC=827.

So link function probit, **gives an appreciably better fit**.

**2) link= c-log log**

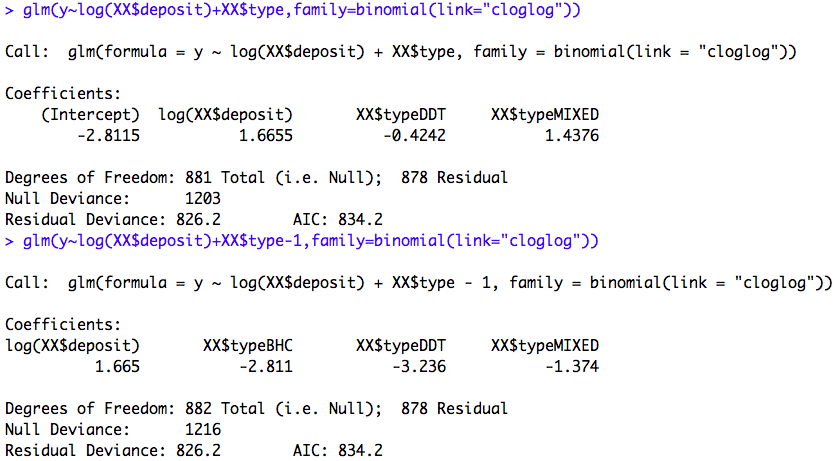


Figure 4.10 R summary table for link=cloglog

When link =cloglog, AIC=834.2 which is larger than link=logit AIC=827.

So link function cloglog, does not give an appreciably better fit.

**Answer to part (e) for the c-log log model**

By r function, beta1= -1.862, beta2= -1.438.

Potency of the combination relative to DDT is exp(1.862)= 6.436597.

Potency of the combination relative to BHC is exp(1.438)= 4.212263.

**90% CI**

90% CI for beta1 (-2.107790, -1.622850)

90% CI for beta2 (-1.664501, -1.216097)

90% CI for potency of the combination relative to DDT is (5.067512, 8.230033)

90% CI for potency of the combination relative to BHC is (3.373993, 5.283036)

(g) Under the linear logistic model, estimate the combination dose required to give a 99% kill rate, and obtain a 90% confidence interval for this dose.

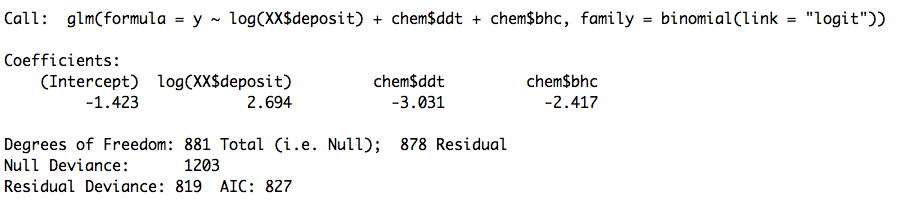


Figure 4.11 Linear Logistic Model



(Intercept) : (-1.891864, -0.9558493)

log(XX$deposit) : (2.348604,3.0552145)

When give a 99% kill rate, the combination Iog(dose)=2.2339, dose=9.3367

90% confidence interval for this log(dose) is (1.816884, 2.76206)

90% confidence interval for this dose is (6.15265, 15.8324)

(h) Give a brief summary of your conclusions regarding the effectiveness of these three insecticides.

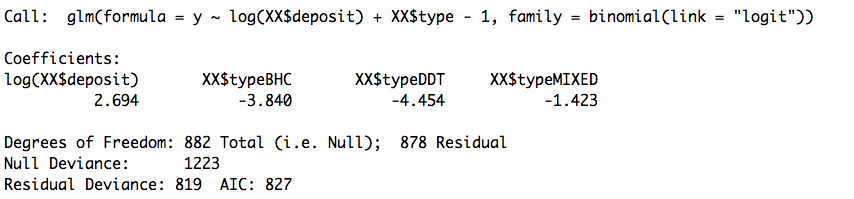


Figure 4.12 R summary table for effectiveness of insecticide

Based on the r summary table, we can find that the mixed insecticide can have a better effect on the kill rate since the coefficient for MIXED term is bigger than the other two. BHC ranks second and DDT has the worst effect on kill rate.

In a word, ranking of the effectiveness of these three insecticides from high to low is: MIXED, BHC, DDT.

**Appendix**

|  |
| --- |
| #1(a)  a=read.table("fungus.txt",header = TRUE)  Y=c(a$X1,a$X2,a$X3,a$X4,a$X5,a$X6,a$X7,a$X8)  temp=c("55", "60", "65", "70", "75", "80", "85","55", "60", "65", "70", "75", "80", "85",  "55", "60", "65", "70", "75", "80", "85","55", "60", "65", "70", "75", "80", "85",  "55", "60", "65", "70", "75", "80", "85","55", "60", "65", "70", "75", "80", "85",  "55", "60", "65", "70", "75", "80", "85","55", "60", "65", "70", "75", "80", "85")  isolate=c("X1","X1","X1","X1","X1","X1","X1","X2","X2","X2","X2","X2","X2","X2","X3","X3","X3","X3","X3","X3","X3",  "X4","X4","X4","X4","X4","X4","X4","X5","X5","X5","X5","X5","X5","X5","X6","X6","X6","X6","X6","X6","X6",  "X7","X7","X7","X7","X7","X7","X7","X8","X8","X8","X8","X8","X8","X8")  lm.reg=lm(Y~isolate+temp)  summary(lm.reg)  anova(lm.reg)  #1(b)  fug = read.table("fungus.txt",header = T)  fug = unlist(fug)[-(1:7)]  fung = data.frame("fungus" = fug,"temp" = factor(rep(seq(55,85,by=5),8)),"iso" = factor(rep(1:8,each = 7)))  fung  temp  a=read.table("fungus.txt",header = TRUE)  a  Y=c(a$X1,a$X2,a$X3,a$X4,a$X5,a$X6,a$X7,a$X8)  x=rep(seq(55,85,by=5),8)  x2=x^2;x3=x^3;x4=x^4;x5=x^5;x6=x^6  lm6=lm(Y~x+x2+x3+x4+x5+x6)  anova(lm6)  lm0=lm(Y~1)  lm1=lm(Y~1+x)  lm2=lm(Y~1+x+x2)  lm3=lm(Y~x+x2+x3)  lm4=lm(Y~x+x2+x3+x4)  lm5=lm(Y~x+x2+x3+x4+x5)  anova(lm0,lm1)  anova(lm1,lm2)  anova(lm2,lm3)  anova(lm3,lm4)  anova(lm4,lm6)  anova(lm3,lm6)  #1(c)  lm3=lm(Y~x+x2+x3)  fun=function(x){  return(-sum(c(1,x,x^2,x^3)\*lm3$coefficients))  }  optimize(fun,c(55,85)) |
| #2(a)  b=read.table("flushot.txt",header = TRUE)  b  glm.reg=glm(data=b,y~x1+x2+x3,family=binomial(link="logit"))  summary(glm.reg)  ci95=confint.default(glm.reg)  #2(b)  round(cbind(exp(glm.reg$coefficients),exp(confint.default(glm.reg))),3)  #2(c)  pnew = predict.glm(glm.reg, newdata=data.frame(x1=55,x2=60,x3=1), se.fit=T, type="link")  hat = pnew$fit  se.hat = pnew$se.fit  phat = exp(hat)/(1+exp(hat))  data.frame("lowerbound"=hat-qnorm(0.975)\*se.hat,"upperbound"=hat+qnorm(0.975)\*se.hat)  uphat = exp(hat-qnorm(0.975)\*se.hat)/(1+exp(hat-qnorm(0.975)\*se.hat))  lohat = exp(hat+qnorm(0.975)\*se.hat)/(1+exp(hat+qnorm(0.975)\*se.hat))  data.frame("lower bound"=uphat ,"upper bound"=lohat) # c.i for estimated prob  #2(e)  anova(glm(data=b,y~x1+x2,family = binomial("logit")),glm.reg,test = "Chisq")  #2(f)  b=read.table("flushot.txt",header = TRUE)  glm.reg=glm(data=b,y~x1+x2+x3,family=binomial(link="logit"))  glm0=glm(data=b,y~1,family = binomial("logit"))  #aic  step(glm0,scope=list(upper=glm.reg),direction = "both")  step(glm0,scope=list(upper=glm.reg),direction = "forward")  step(glm.reg)  #bic  step(glm0,scope=list(upper=glm.reg),direction = "both",k=log(length(y)))  step(glm0,scope=list(upper=glm.reg),direction = "forward",k=log(length(y)))  step(glm.reg,k=log(length(y)))  #final model  glm.reg=glm(data=b,y~x1+x2,family=binomial(link="logit"))  glm.reg  #2(g)  b=read.table("flushot.txt",header = TRUE)  glm.reg.final=glm(y~x1+x2,family=binomial(link="logit"))  install.packages("pROC")  library(pROC)  y=b$y  flushot.roc=roc(y~fitted(glm.reg.final))  plot(flushot.roc,cex.lab=1.5)  auc(flushot.roc) |
| #4(a)  par(mfrow=c(1,1))  kill.rate=c(3/50,5/49,19/47,19/38,24/49,35/50,2/50,14/49,20/50,27/50,41/50,40/50,28/50,37/50,46/50,48/50,48/50,50/50)  doses=c(2.00,2.64,3.48,4.59,6.06,8.00,2.00,2.64,3.48,4.59,6.06,8.00,2.00,2.64,3.48,4.59,6.06,8.00)  log(doses)  doses.killrate=lm(kill.rate~doses)  plot(log(doses),kill.rate,cex.lab=1.5)  #4(b)  data=data.frame("deposit"=c(2.00,2.64,3.48,4.59,6.06,8.00),  "ddt"=c(3/50,5/49,19/47,19/38,24/49,35/50),  "bhc"=c(2/50,14/49,20/50,27/50,41/50,40/50),  "both"=c(28/50,37/50,46/50,48/50,48/50,50/50))  logit = function(x){  log(x/(1-x))  }  data$deposit = log(data$deposit)  plot(data$deposit,logit(data$ddt),ylim=c(-4,4),xlab="deposit",ylab="logit(kill rate)",cex.lab=1.5)  points(data$deposit,logit(data$bhc),col="red")  points(data$deposit,logit(data$both),col="blue")  d = lm(logit(ddt)~deposit,data=data)  abline(reg=d)  b = lm(logit(bhc)~deposit,data=data)  abline(reg=b)  combine = lm(logit(data$both[1:5])~deposit[1:5],data=data)  abline(reg=combine)  #4(c)  insect=c(rep(0,6),rep(1,6),rep(2,6))  insc\_type = factor(insect, levels = c(0, 1, 2),  labels = c("DDT", "BHC", "MIXED"))  dose=rep(c(2.00, 2.64, 3.48, 4.59, 6.06, 8.00),3)  killed=c(3,5,19,19,24,35,2,14,20,27,41,40,28,37,46,48,48,50)  unkilled=c(50,49,47,38,49,50,50,49,rep(50,10))-killed  data=data.frame("killed"=killed,"unkilled"=unkilled,"deposit"=dose,"type"=insc\_type)  # making binary var  y=c()  for(i in 1:length(killed)){  yi=c(rep(1,killed[i]),rep(0,unkilled[i]))  y=c(y,yi)  }  total=killed+unkilled  cumsum(total)  total=c(0,total)  XX = matrix(0,ncol = 2,nrow =sum(total))  X=as.matrix(data[,c(3,4)])  cum=cumsum(total)  for(i in 1:(length(cum)-1)){  k=((cum[i]+1):(cum[i+1]))  v=X[i,]  for(j in 1:length(k)){  XX[k[j],]=v  }  }  XX=data.frame(XX)  XX$X1=as.numeric(as.character(XX$X1))  colnames(XX)=c("deposit","type")  head(XX)  ddt=c(rep(1,283),rep(0,299),rep(0,300))  bhc=c(rep(0,283),rep(1,299),rep(0,300))  chem=data.frame("ddt"=ddt,"bhc"=bhc)  parallel= glm(y~log(XX$deposit)+chem$ddt+chem$bhc+log(XX$deposit)\*chem$ddt+log(XX$deposit)\*chem$bhc,family=binomial(link="logit"))  summary(parallel)  #4(d)  #regression  glm1=glm(y~log(XX$deposit)+XX$type,family=binomial(link="logit"))  glm2=glm(y~log(XX$deposit)+XX$type-1,family=binomial(link="logit"))  #var-cor  summary(glm1)$cov.unscaled  summary(glm2)$cov.unscaled  #4(e)  confint(glm2,level = 0.9)  #4(f)  glm(y~log(XX$deposit)+XX$type,family=binomial(link="probit"))  glm(y~log(XX$deposit)+XX$type-1,family=binomial(link="probit"))  glm(y~log(XX$deposit)+XX$type,family=binomial(link="cloglog"))  glmc=glm(y~log(XX$deposit)+XX$type-1,family=binomial(link="cloglog"))  confint(glmc,level = 0.9) |