Exam 3 - take home

Vinay Vaida

2023-12-05

# 1. Maximum Likelihood Estimation

(10 pts)

The MLE for estimating the probability in a null model.

In the notes I have explained how the parameters for the logistic model (the ’s) are obtained by maximizing the log likelihood function. We are going to do the same but for a simpler model.

1. Write down the log likelihood for the null model, that is, the model where all probabilities are the same: .

Follow the notes and replace the ’s by . Answer:

L = Ans: a) The log likelihood for the null model, where all probabilities are the same (), is given by:

In this case, for binary outcomes (0 or 1), we can express as:

Substitute this expression into the log likelihood:

Now, since , we can express and in terms of :

1. Find , the log likelihood function and find it’s derivative with respect to .

Find the MLE estimator for as the place where the maximum is achieved.

Hint: follow the notes. Answer:

To find the derivative with respect to , we take the derivative of the above expression:

Setting this derivative to zero gives the MLE estimator for :

If you are at a loss with this problem, ask for help.

# 2. Logistic Regression

(20 pts)

#### Titanic dataset:

Variables:

* PassengerId: Passenger ID
* Survived: Passenger Survival Indicator
* Pclass: Passenger Class
* Name: Name
* Sex: Sex
* Age: Age
* SibSp: Number of Siblings/Spouses Aboard
* Parch: Number of Parents/Children Aboard
* Ticket: Ticket Number
* Fare: Passenger Fare
* Cabin: Cabin
* Embarked: Port of Embarkation

## Load the datasets  
library(titanic)

## Warning: package 'titanic' was built under R version 4.3.2

str(titanic\_train)

## 'data.frame': 891 obs. of 12 variables:  
## $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Survived : int 0 1 1 1 0 0 0 0 1 1 ...  
## $ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...  
## $ Name : chr "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...  
## $ Sex : chr "male" "female" "female" "female" ...  
## $ Age : num 22 38 26 35 35 NA 54 2 27 14 ...  
## $ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...  
## $ Parch : int 0 0 0 0 0 0 0 1 2 0 ...  
## $ Ticket : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Cabin : chr "" "C85" "" "C123" ...  
## $ Embarked : chr "S" "C" "S" "S" ...

This library has 2 datasets, titanic\_train and titanic\_test. We will work with titanic\_train

sum(is.na(titanic\_train))

## [1] 177

mytitanic<-titanic\_train[complete.cases(titanic\_train),]  
n<-dim(mytitanic)[1]  
set.seed(1)  
indx<-sample(1:n,.8\*n)  
titanic\_mytrain<-mytitanic[indx,]  
titanic\_mytest<-mytitanic[-indx,]

Create a logistic model with no interactions. Be careful about the variables you include.

titanicmod1<-glm(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked,   
 data = titanic\_mytrain,   
 family = "binomial")  
summary(titanicmod1)

##   
## Call:  
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +   
## Fare + Embarked, family = "binomial", data = titanic\_mytrain)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 18.240277 608.034234 0.030 0.97607   
## Pclass -1.304993 0.203151 -6.424 1.33e-10 \*\*\*  
## Sexmale -2.875194 0.260497 -11.037 < 2e-16 \*\*\*  
## Age -0.044061 0.009261 -4.758 1.96e-06 \*\*\*  
## SibSp -0.406014 0.156637 -2.592 0.00954 \*\*   
## Parch -0.051546 0.143472 -0.359 0.71939   
## Fare -0.001061 0.004281 -0.248 0.80421   
## EmbarkedC -12.055925 608.033855 -0.020 0.98418   
## EmbarkedQ -12.825533 608.034093 -0.021 0.98317   
## EmbarkedS -12.573055 608.033831 -0.021 0.98350   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 769.86 on 570 degrees of freedom  
## Residual deviance: 480.25 on 561 degrees of freedom  
## AIC: 500.25  
##   
## Number of Fisher Scoring iterations: 13

1. (4pts) For this section, you need to know what the variables are.

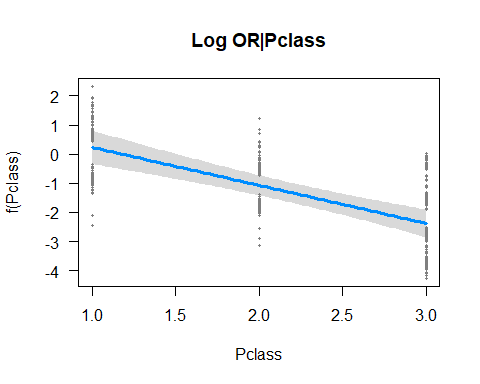
* Interpret the coefficient for Pclass-
  + The coefficient for Pclass is -1.304993.
  + The interpretation is that for a one-unit increase in Pclass (moving from a lower class to a higher class), the log-odds of survival decrease by approximately 1.30.
  + The negative sign indicates that a higher passenger class is associated with a lower likelihood of survival.
* Interpret the coefficient for Sex -
  + The coefficient for Sexmale is -2.875194.
  + The interpretation is that being male (compared to being female) is associated with a decrease in the log-odds of survival by approximately 2.88.
  + The negative sign indicates that males are less likely to survive compared to females.
* Interpret the coefficient for Age -
  + The coefficient for Age is -0.044061.
  + The interpretation is that for a one-unit increase in age, the log-odds of survival decrease by approximately 0.044.
  + The negative sign indicates that older passengers are less likely to survive.

Visualization of fitted values vs variables in model.

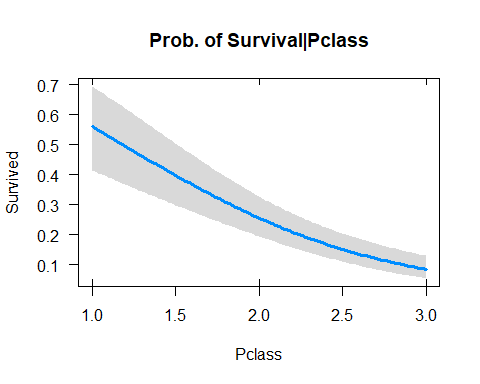
library(visreg)

## Warning: package 'visreg' was built under R version 4.3.2

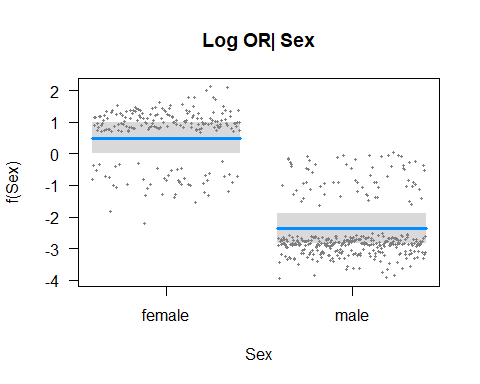
visreg(titanicmod1, "Pclass", partial = TRUE,main="Log OR|Pclass") #plots the log odds ratios



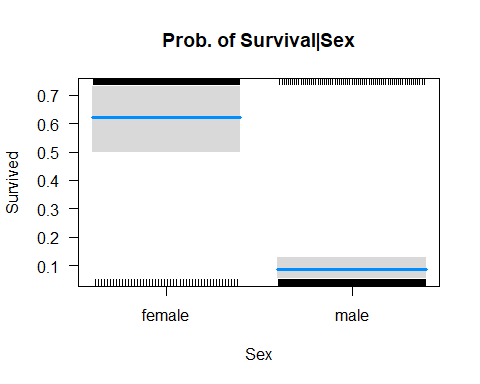
visreg(titanicmod1, "Pclass", scale="response", partial=FALSE,main="Prob. of Survival|Pclass") #plots probabilities



visreg(titanicmod1, "Sex", partial = TRUE,main="Log OR| Sex")

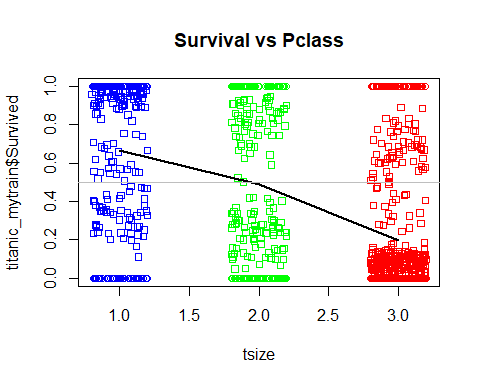


visreg(titanicmod1, "Sex",scale="response", partial=FALSE,main="Prob. of Survival|Sex") #plots probabilities

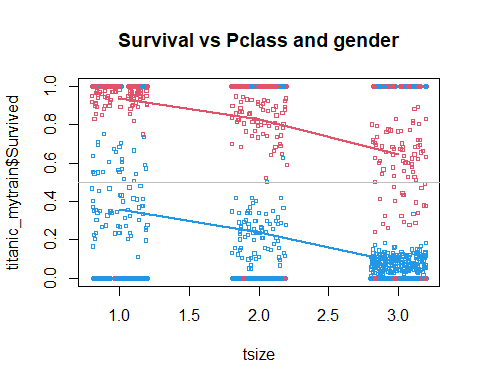


Some more visualizations of fitted vs Pclass:

x<-jitter(titanic\_mytrain$Pclass)  
plot(titanic\_mytrain$Survived~x,col=c("blue","green","red")[titanic\_mytrain$Pclass],xlab="tsize",main="Survival vs Pclass")  
points(x,titanicmod1$fitted,col=c("blue","green","red")[titanic\_mytrain$Pclass],pch=22)  
#add fitted values curve  
lines(lowess(titanic\_mytrain$Pclass,titanicmod1$fitted),col="black",lwd=2)  
#look at a threshold of .5  
abline(h=.5,col='grey')



x<-jitter(titanic\_mytrain$Pclass)  
scol<-ifelse(titanic\_mytrain$Sex=="female",2,4)  
plot(titanic\_mytrain$Survived~x,col=scol,xlab="tsize",  
 pch=19,cex=.7,main="Survival vs Pclass and gender")  
points(x,titanicmod1$fitted,col=scol,pch=22,cex=.5)  
#add fitted values curve  
lines(lowess(titanic\_mytrain$Pclass[titanic\_mytrain$Sex=="female"],titanicmod1$fitted[titanic\_mytrain$Sex=="female"]),col=2,lwd=2,cex=.5)  
lines(lowess(titanic\_mytrain$Pclass[titanic\_mytrain$Sex=="male"],titanicmod1$fitted[titanic\_mytrain$Sex=="male"]),col=4,lwd=2,cex=.5)  
#look at a threshold of .5  
abline(h=.5,col='grey')

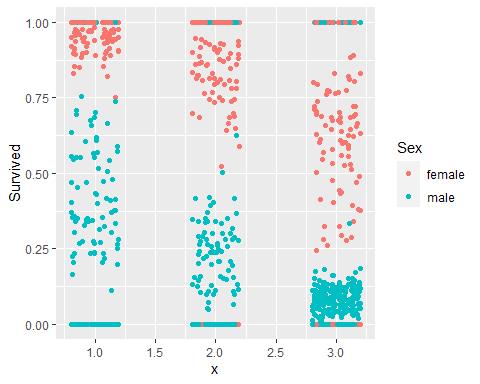


Same thing with ggplot

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.3.2

df<-titanic\_mytrain[,c(2,5)]  
df$x<-x  
ggplot(df,aes(x=x,y=Survived,col=Sex))+geom\_point()+  
 geom\_point(aes(x=x,y=titanicmod1$fitted.values,col=Sex))



1. (2pts) Create a logistic model with interactions.

titanicmod2<-glm(Survived ~ Pclass \* Sex + Age + SibSp + Parch + Fare + Embarked,   
 data = titanic\_mytrain,   
 family = "binomial")  
  
summary(titanicmod2)

##   
## Call:  
## glm(formula = Survived ~ Pclass \* Sex + Age + SibSp + Parch +   
## Fare + Embarked, family = "binomial", data = titanic\_mytrain)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 19.978079 607.714749 0.033 0.973775   
## Pclass -2.640753 0.437895 -6.031 1.63e-09 \*\*\*  
## Sexmale -6.799641 1.154750 -5.888 3.90e-09 \*\*\*  
## Age -0.046143 0.009772 -4.722 2.33e-06 \*\*\*  
## SibSp -0.360349 0.154424 -2.334 0.019621 \*   
## Parch 0.022900 0.152056 0.151 0.880287   
## Fare -0.004645 0.005134 -0.905 0.365611   
## EmbarkedC -10.212718 607.714427 -0.017 0.986592   
## EmbarkedQ -10.837871 607.714659 -0.018 0.985771   
## EmbarkedS -10.843780 607.714375 -0.018 0.985764   
## Pclass:Sexmale 1.609252 0.430808 3.735 0.000187 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 769.86 on 570 degrees of freedom  
## Residual deviance: 462.28 on 560 degrees of freedom  
## AIC: 484.28  
##   
## Number of Fisher Scoring iterations: 13

b2) (1pts) Do backwards selection, use trace = 0

titanicmod3<-step(titanicmod2, direction = "backward", trace = 0)  
summary(titanicmod3)

##   
## Call:  
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Pclass:Sex,   
## family = "binomial", data = titanic\_mytrain)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 8.879950 1.200533 7.397 1.40e-13 \*\*\*  
## Pclass -2.533710 0.400352 -6.329 2.47e-10 \*\*\*  
## Sexmale -6.589890 1.136142 -5.800 6.62e-09 \*\*\*  
## Age -0.046168 0.009708 -4.756 1.98e-06 \*\*\*  
## SibSp -0.393931 0.144457 -2.727 0.006392 \*\*   
## Pclass:Sexmale 1.533478 0.422660 3.628 0.000285 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 769.86 on 570 degrees of freedom  
## Residual deviance: 466.70 on 565 degrees of freedom  
## AIC: 478.7  
##   
## Number of Fisher Scoring iterations: 6

b3) (4pts) Interpret all interaction terms coefficients.

Pclass:Sex (Interaction):

Estimate: 1.533478 Interpretation: This represents the additional change in the log-odds of survival for the interaction effect between Pclass and Sex. It shows how the effect of Pclass on the log-odds of survival differs between the levels of Sex.

1. (3pts) Compare titanicmod1,titanicmod3 (using anova)

anova\_output<-anova(titanicmod1, titanicmod3, test = "Chi")  
print(anova\_output)

## Analysis of Deviance Table  
##   
## Model 1: Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked  
## Model 2: Survived ~ Pclass + Sex + Age + SibSp + Pclass:Sex  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 561 480.25   
## 2 565 466.70 -4 13.551

Write the null and alternative hypothesis corresponding to this test:

Null hypothesis:

Alternative hypothesis:

p-value:

Conclusion:

In the context of an analysis of deviance table, the null and alternative hypotheses are typically formulated as follows:

**Null hypothesis (H0):** - The null hypothesis asserts that the simpler model (Model 2) is just as good as the more complex model (Model 1). In terms of deviance, it means that the additional interaction term (Pclass:Sex) in Model 1 does not significantly improve the model fit compared to the reduced model (Model 2).

**Alternative hypothesis (H1):** - The alternative hypothesis suggests that the more complex model (Model 1) is a better fit than the simpler model (Model 2). In terms of deviance, it implies that the additional interaction term (Pclass:Sex) in Model 1 contributes significantly to the model fit.

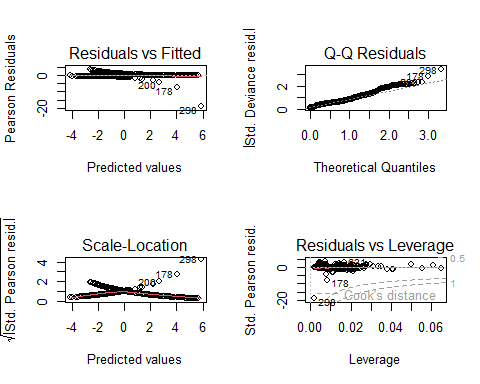
**p-value:** - The p-value (given as 13.551) is the probability of observing a test statistic as extreme as the one calculated (or more extreme) under the assumption that the null hypothesis is true.

**Conclusion:** - If the p-value is greater than the significance level (commonly set at 0.05), you would fail to reject the null hypothesis. This would suggest that there is not enough evidence to conclude that the additional interaction term in Model 1 significantly improves the model fit compared to the simpler Model 2. - If the p-value is less than the significance level, you would reject the null hypothesis. This would suggest that the additional interaction term in Model 1 contributes significantly to the model fit, and Model 1 is preferred over the simpler Model 2.

In this case, the p-value is not explicitly provided, but it is given as 13.551. Typically, a p-value larger than 0.05 would lead to a failure to reject the null hypothesis. However, without the exact p-value, the conclusion cannot be determined definitively. If the p-value is less than 0.05, it would suggest evidence against the null hypothesis, favoring Model 1.

1. (1pt) Plot the assessment plots for titanicmod3

par(mfrow=c(2,2)) # Setting up a 2x2 layout for the plots  
plot(titanicmod3)



### Confusion matrix and statistics

1. (1pt) Compute the confusion matrix for both models on the testing data.

library(caret)

## Loading required package: lattice

# for first model  
fit\_test<-predict(titanicmod1,newdata=titanic\_mytest[,-2],type="response")  
cm1<-confusionMatrix(data=as.factor((fit\_test>.5)\*1), reference=as.factor(titanic\_mytest$Survived))  
cm1$table

## Reference  
## Prediction 0 1  
## 0 69 23  
## 1 14 37

# for second model  
fit\_test<-predict(titanicmod3, newdata = titanic\_mytest[,-2], type = "response")  
cm3 <- confusionMatrix(data = as.factor((fit\_test > 0.5) \* 1), reference = as.factor(titanic\_mytest$Survived))  
cm3$table

## Reference  
## Prediction 0 1  
## 0 75 24  
## 1 8 36

Which of the two models performs better?

accuracy for titanicmod1 model=(69+37)/(69+23+14+37)= 0.74 accuracy for titanicmod3 model=(75+36)/(75+24+8+36)= 0.77 The model with a higher accuracy is considered better. Hence model “titanicmod3” is better model than model “titanicmod1”

1. (2pt) Compute the auc for both models

library(Epi)

## Warning: package 'Epi' was built under R version 4.3.2

# titanic1  
library(pROC)

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

# For model 1  
roc1 <- roc(titanic\_mytest$Survived, predict(titanicmod1, newdata = titanic\_mytest[,-2]))

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

# Display AUC values  
paste("AUC for Model 1:", round(auc(roc1), 4))

## [1] "AUC for Model 1: 0.7944"

# For model 3  
roc3 <- roc(titanic\_mytest$Survived, predict(titanicmod3, newdata = titanic\_mytest[,-2]))

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

paste("AUC for Model 3:", round(auc(roc3), 4))

## [1] "AUC for Model 3: 0.8024"

#titanic3

AUC for titanic1:0.7944

AUC for titanic3: 0.8024

1. (2pts) Write a short comment comparing the two models.

Model 1 has an AUC of 0.7944. Model 3 has a slightly higher AUC of 0.8024. A higher AUC generally indicates better discrimination. Therefore, based on the AUC values: Model 3 performs slightly better in terms of discriminatory power compared to Model 1.

True Positives (TP): Model 3 has more true positives (36) compared to Model 1 (37). False Positives (FP): Model 3 has fewer false positives (8) compared to Model 1 (14). True Negatives (TN): Model 3 has more true negatives (75) compared to Model 1 (69). False Negatives (FN): Model 1 has fewer false negatives (23) compared to Model 3 (24). In summary, Model 3 generally performs better than Model 1 in terms of true positives and true negatives, indicating improved classification accuracy.

# 3. Breast Cancer

(40 pts) Variables: \* age and agegp, \* menopause, \* tsize and tumorsize” \* invnodes2 and inv.nodes \* node.caps \* deg.malig: severity of malignity  
\* breast (left/right) and breastquad (L/R and up/central/low) \* irradiate \* Class and Y.

We will use Y as the response variable

breast <- read.csv("C:/Users/vinay/OneDrive/Documents/Applied Stats/Assignment 6/breast.csv")  
  
str(breast)

## 'data.frame': 286 obs. of 14 variables:  
## $ age : int 45 55 55 45 45 55 55 45 45 45 ...  
## $ agegp : chr "40-49" "50-59" "50-59" "40-49" ...  
## $ menopause : chr "premeno" "ge40" "ge40" "premeno" ...  
## $ tsize : int 17 17 37 37 32 27 42 12 2 42 ...  
## $ tumorsize : chr "15-19" "15-19" "35-39" "35-39" ...  
## $ invnodes2 : int 1 1 1 1 4 4 1 1 1 16 ...  
## $ inv.nodes : chr "0-2" "0-2" "0-2" "0-2" ...  
## $ node.caps : chr "yes" "no" "no" "yes" ...  
## $ deg.malig : int 3 1 2 3 2 2 3 2 2 2 ...  
## $ breast : chr "right" "right" "left" "right" ...  
## $ breastquad: chr "left\_up" "central" "left\_low" "left\_low" ...  
## $ irradiate : chr "no" "no" "no" "yes" ...  
## $ Class : chr "recurrence-events" "no-recurrence-events" "recurrence-events" "no-recurrence-events" ...  
## $ Y : int 1 0 1 0 1 0 0 0 0 0 ...

Missing data have been coded with ?. Change it ti NA and then remove the NA’s.

breast[breast == "?"] <- NA # Replace "?" with NA  
# Remove rows with NA values  
newbreast <-na.omit(breast ) # removes all rows with missing data NA

1. (2pt) Develop a logistic model with tsize and deg.malig

logmod<-glm(Y ~ tsize + deg.malig, data = newbreast, family = "binomial")  
summary(logmod)

##   
## Call:  
## glm(formula = Y ~ tsize + deg.malig, family = "binomial", data = newbreast)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.99098 0.61564 -6.483 9.01e-11 \*\*\*  
## tsize 0.03080 0.01407 2.189 0.0286 \*   
## deg.malig 1.03937 0.21402 4.856 1.20e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 334.78 on 276 degrees of freedom  
## Residual deviance: 297.96 on 274 degrees of freedom  
## AIC: 303.96  
##   
## Number of Fisher Scoring iterations: 4

1. (4pts) Develop a logistic model with all variables. Use age but not agegp, tsize but not tumorsize, and invnodes2 but not inv.nodes.

logmod2<-glm(Y ~ age + menopause + tsize + invnodes2 + node.caps + deg.malig + breast + breastquad + irradiate,   
 data = newbreast,   
 family = "binomial")  
summary(logmod2)

##   
## Call:  
## glm(formula = Y ~ age + menopause + tsize + invnodes2 + node.caps +   
## deg.malig + breast + breastquad + irradiate, family = "binomial",   
## data = newbreast)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.11393 1.49449 -2.084 0.037196 \*   
## age -0.01430 0.02067 -0.692 0.489100   
## menopauselt40 -15.13643 1031.38329 -0.015 0.988291   
## menopausepremeno 0.14789 0.42084 0.351 0.725276   
## tsize 0.02179 0.01526 1.428 0.153357   
## invnodes2 0.04757 0.04177 1.139 0.254822   
## node.capsyes 0.49779 0.39812 1.250 0.211167   
## deg.malig 0.85750 0.23832 3.598 0.000321 \*\*\*  
## breastright -0.20804 0.32133 -0.647 0.517348   
## breastquadleft\_low 0.20002 0.65218 0.307 0.759070   
## breastquadleft\_up -0.01243 0.66507 -0.019 0.985092   
## breastquadright\_low -0.20204 0.84091 -0.240 0.810129   
## breastquadright\_up 0.60336 0.73885 0.817 0.414144   
## irradiateyes 0.50774 0.34536 1.470 0.141510   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 334.78 on 276 degrees of freedom  
## Residual deviance: 279.24 on 263 degrees of freedom  
## AIC: 307.24  
##   
## Number of Fisher Scoring iterations: 15

1. (1pt) Use backwards selection to clean the model.

logmod3<-step(logmod2, direction = "backward")

## Start: AIC=307.24  
## Y ~ age + menopause + tsize + invnodes2 + node.caps + deg.malig +   
## breast + breastquad + irradiate  
##   
## Df Deviance AIC  
## - breastquad 4 281.34 301.34  
## - breast 1 279.66 305.66  
## - menopause 2 281.70 305.70  
## - age 1 279.72 305.72  
## - invnodes2 1 280.63 306.63  
## - node.caps 1 280.77 306.77  
## <none> 279.24 307.24  
## - tsize 1 281.28 307.28  
## - irradiate 1 281.36 307.36  
## - deg.malig 1 293.15 319.15  
##   
## Step: AIC=301.34  
## Y ~ age + menopause + tsize + invnodes2 + node.caps + deg.malig +   
## breast + irradiate  
##   
## Df Deviance AIC  
## - age 1 281.69 299.69  
## - breast 1 281.69 299.69  
## - menopause 2 284.05 300.05  
## - invnodes2 1 282.48 300.48  
## - irradiate 1 283.10 301.10  
## - node.caps 1 283.18 301.18  
## <none> 281.34 301.34  
## - tsize 1 284.28 302.28  
## - deg.malig 1 295.59 313.59  
##   
## Step: AIC=299.69  
## Y ~ menopause + tsize + invnodes2 + node.caps + deg.malig + breast +   
## irradiate  
##   
## Df Deviance AIC  
## - breast 1 282.09 298.09  
## - invnodes2 1 282.98 298.98  
## - node.caps 1 283.45 299.45  
## - irradiate 1 283.48 299.48  
## <none> 281.69 299.69  
## - menopause 2 285.75 299.75  
## - tsize 1 284.53 300.53  
## - deg.malig 1 296.44 312.44  
##   
## Step: AIC=298.09  
## Y ~ menopause + tsize + invnodes2 + node.caps + deg.malig + irradiate  
##   
## Df Deviance AIC  
## - invnodes2 1 283.38 297.38  
## - irradiate 1 283.79 297.79  
## - node.caps 1 283.85 297.85  
## <none> 282.09 298.09  
## - menopause 2 286.15 298.15  
## - tsize 1 284.81 298.81  
## - deg.malig 1 297.25 311.25  
##   
## Step: AIC=297.38  
## Y ~ menopause + tsize + node.caps + deg.malig + irradiate  
##   
## Df Deviance AIC  
## <none> 283.38 297.38  
## - menopause 2 287.49 297.49  
## - irradiate 1 285.62 297.62  
## - tsize 1 286.05 298.05  
## - node.caps 1 287.82 299.82  
## - deg.malig 1 300.87 312.87

summary(logmod3)

##   
## Call:  
## glm(formula = Y ~ menopause + tsize + node.caps + deg.malig +   
## irradiate, family = "binomial", data = newbreast)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.06335 0.69482 -5.848 4.97e-09 \*\*\*  
## menopauselt40 -15.09420 992.24518 -0.015 0.9879   
## menopausepremeno 0.36792 0.29887 1.231 0.2183   
## tsize 0.02387 0.01468 1.626 0.1041   
## node.capsyes 0.73576 0.34819 2.113 0.0346 \*   
## deg.malig 0.93116 0.23266 4.002 6.27e-05 \*\*\*  
## irradiateyes 0.50555 0.33547 1.507 0.1318   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 334.78 on 276 degrees of freedom  
## Residual deviance: 283.38 on 270 degrees of freedom  
## AIC: 297.38  
##   
## Number of Fisher Scoring iterations: 15

1. (4pts) Compare the models using anova.

# Fit the models  
logmod <- glm(Y ~ tsize + deg.malig, data = newbreast, family = "binomial")  
logmod\_all <- glm(Y ~ age + menopause + tsize + invnodes2 + node.caps + deg.malig + breast + breastquad + irradiate,  
 data = newbreast, family = "binomial")  
logmod3 <- glm(Y ~ menopause + tsize + node.caps + deg.malig + irradiate, data = newbreast, family = "binomial")  
  
# Perform ANOVA  
anova\_result <- anova(logmod, logmod\_all, logmod3, test = "Chisq")  
print(anova\_result)

## Analysis of Deviance Table  
##   
## Model 1: Y ~ tsize + deg.malig  
## Model 2: Y ~ age + menopause + tsize + invnodes2 + node.caps + deg.malig +   
## breast + breastquad + irradiate  
## Model 3: Y ~ menopause + tsize + node.caps + deg.malig + irradiate  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 274 297.96   
## 2 263 279.24 11 18.7281 0.06615 .  
## 3 270 283.38 -7 -4.1439 0.76307   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

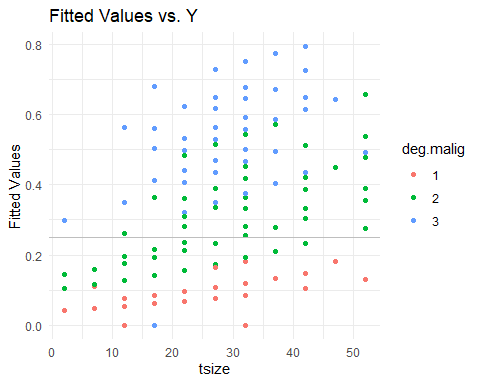
Write the null and alternative hypothesis corresponding to this test: Null hypothesis: The models (Model 1 and Model 2, as well as Model 1 and Model 3) are equivalent, and the additional predictors in the more complex models (Model 2 and Model 3) do not significantly contribute to explaining the variability in the response variable Y.

Alternative hypothesis: There is a significant difference in deviance between the simpler model (Model 1) and the more complex models (Model 2 and Model 3), indicating that the additional predictors in the complex models contribute significantly to explaining the variability in the response variable Y.

p-value: The p-values associated with the tests are 0.06615 (Model 1 vs. Model 2) and 0.76307 (Model 1 vs. Model 3).

1. (4pts) Plot the fitted values of logmod3 over a plot of Y vs tsize. Color according to deg.malig. Add a horizontal line at y=.25. If using ggplot: geom\_hline(yintercept = .22,col=“grey”)

library(ggplot2)  
  
# Plotting the fitted values  
plot\_data <- data.frame(tsize = newbreast$tsize, fitted\_values = predict(logmod3, type = "response"))  
plot\_data$deg.malig <- as.factor(newbreast$deg.malig)  
  
ggplot(plot\_data, aes(x = tsize, y = fitted\_values, color = deg.malig)) +  
 geom\_point() +  
 geom\_hline(yintercept = 0.25, col = "grey") +  
 labs(title = "Fitted Values vs. Y",  
 x = "tsize",  
 y = "Fitted Values") +  
 theme\_minimal()



1. (4pts) Use the variables in the logmod3 but now add a polynomial of degree 2 on tsize and deg.malig.

# Create a new variable for the squared term of tsize  
newbreast$tsize\_sq <- newbreast$tsize^2  
  
# Create a new variable for the squared term of deg.malig  
newbreast$deg\_malig\_sq <- newbreast$deg.malig^2  
  
# Logistic model with polynomial terms  
logmod4 <- glm(Y ~ menopause + tsize + tsize\_sq + node.caps + deg.malig + deg\_malig\_sq + irradiate,   
 family = "binomial",   
 data = newbreast)  
  
# Display the summary  
summary(logmod4)

##   
## Call:  
## glm(formula = Y ~ menopause + tsize + tsize\_sq + node.caps +   
## deg.malig + deg\_malig\_sq + irradiate, family = "binomial",   
## data = newbreast)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.790766 1.641202 -1.700 0.0890 .  
## menopauselt40 -15.207305 984.111346 -0.015 0.9877   
## menopausepremeno 0.407365 0.307931 1.323 0.1859   
## tsize 0.124795 0.070186 1.778 0.0754 .  
## tsize\_sq -0.001733 0.001161 -1.492 0.1357   
## node.capsyes 0.769094 0.359369 2.140 0.0323 \*  
## deg.malig -1.752482 1.361765 -1.287 0.1981   
## deg\_malig\_sq 0.616284 0.319057 1.932 0.0534 .  
## irradiateyes 0.579789 0.345523 1.678 0.0933 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 334.78 on 276 degrees of freedom  
## Residual deviance: 276.86 on 268 degrees of freedom  
## AIC: 294.86  
##   
## Number of Fisher Scoring iterations: 15

1. (4pts) Compare logmod3 and logmod4 using anova

# Compare logmod3 and logmod4 using ANOVA  
anova\_result <- anova(logmod3, logmod4, test = "Chi")  
  
# Display the result  
print(anova\_result)

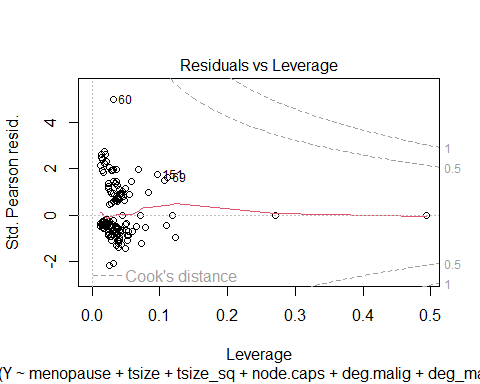
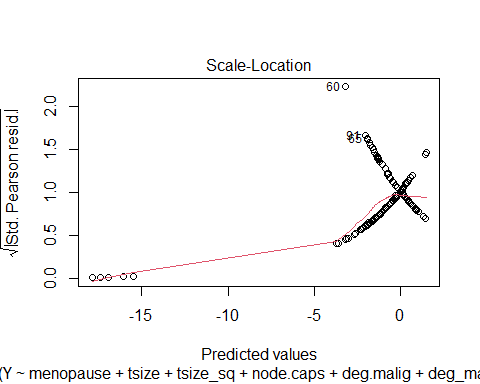
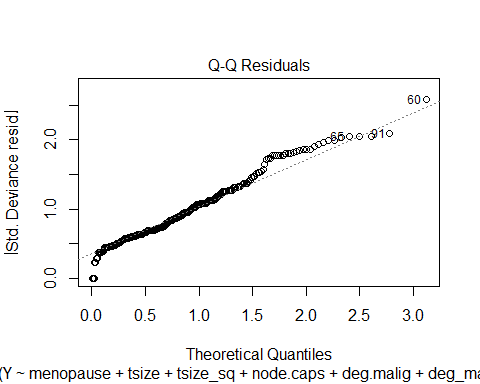
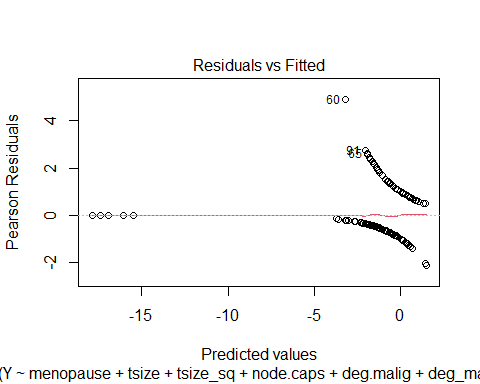
## Analysis of Deviance Table  
##   
## Model 1: Y ~ menopause + tsize + node.caps + deg.malig + irradiate  
## Model 2: Y ~ menopause + tsize + tsize\_sq + node.caps + deg.malig + deg\_malig\_sq +   
## irradiate  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 270 283.38   
## 2 268 276.86 2 6.5193 0.0384 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Which model is preffered?

– The results indicate that Model 2 with the polynomial terms ( and ) has a lower residual deviance compared to Model 1, and the difference in deviance is statistically significant (). This suggests that Model 2 is preferred over Model 1. In conclusion, the addition of the polynomial terms improves the model fit, and Model 2 is preferred for predicting.

1. (1pt) Plot the assessment plots for logmod4

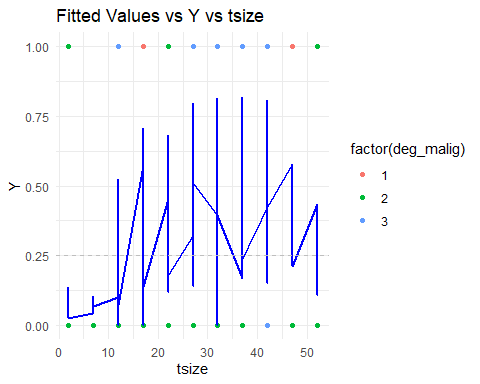
# Plot assessment plots for logmod4  
plot(logmod4, which = c(1, 2, 3, 5))



1. (4pts) Plot the fitted values over a plot of Y vs tsize. Color according to deg.malig. Add a horizontal line at y=.25. If using ggplot: geom\_hline(yintercept = .25,col=“grey”)

# Predicted values from logmod4  
predicted\_probs <- predict(logmod4, type = "response")  
  
# Create a data frame for plotting  
plot\_data <- data.frame(Y = newbreast$Y, tsize = newbreast$tsize, deg\_malig = newbreast$deg.malig, Predicted = predicted\_probs)  
  
# Plot the fitted values over Y vs tsize  
library(ggplot2)  
ggplot(plot\_data, aes(x = tsize, y = Y, color = factor(deg\_malig))) +  
 geom\_point() +  
 geom\_line(aes(y = Predicted), color = "blue", size = 1) +  
 geom\_hline(yintercept = 0.25, linetype = "dashed", color = "grey") +  
 labs(title = "Fitted Values vs Y vs tsize",  
 x = "tsize",  
 y = "Y") +  
 theme\_minimal()

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



1. (6pts) Use Epi::ROC to plot the ROC curve and select a threshold for logmod3 and logmod4. Compute the auc for both models. Note: instead of poly(tsize,2) use tsize +I(tsize^2), and do the same for deg.malig.

auc for logmod3: Suggested threshold:

auc for logmod4: Suggested threshold:

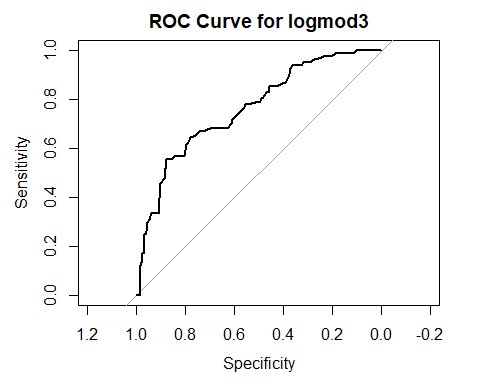
Which model is better?

# Load the Epi package  
library(Epi)  
  
# Create a new variable for the squared term of tsize and deg.malig  
newbreast$tsize\_sq <- newbreast$tsize^2  
newbreast$deg\_malig\_sq <- newbreast$deg.malig^2  
  
# Logistic model with polynomial terms for logmod3  
logmod3 <- glm(Y ~ menopause + tsize + node.caps + deg.malig + irradiate,   
 family = "binomial",   
 data = newbreast)  
  
# Logistic model with polynomial terms for logmod4  
logmod4 <- glm(Y ~ menopause + tsize + tsize\_sq + node.caps + deg.malig + deg\_malig\_sq +   
 irradiate,   
 family = "binomial",   
 data = newbreast)  
  
# ROC curve for logmod3  
roc\_mod3 <- roc(newbreast$Y, predict(logmod3, type = "response"))

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

plot(roc\_mod3, main = "ROC Curve for logmod3")



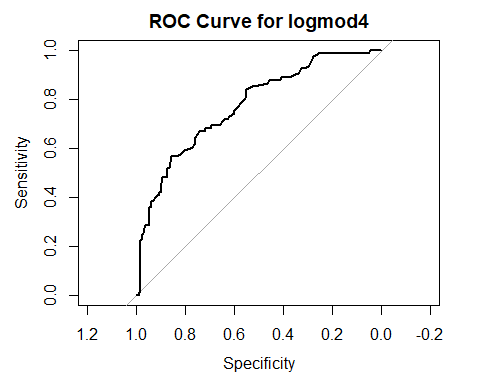
auc\_mod3 <- auc(roc\_mod3)  
suggested\_threshold\_mod3 <- coords(roc\_mod3, "best")$threshold  
cat("AUC for logmod3:", auc\_mod3, "\nSuggested threshold:", suggested\_threshold\_mod3, "\n")

## AUC for logmod3: 0.7616213   
## Suggested threshold: 0.4347756

# ROC curve for logmod4  
roc\_mod4 <- roc(newbreast$Y, predict(logmod4, type = "response"))

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

plot(roc\_mod4, main = "ROC Curve for logmod4")



auc\_mod4 <- auc(roc\_mod4)  
suggested\_threshold\_mod4 <- coords(roc\_mod4, "best")$threshold  
cat("AUC for logmod4:", auc\_mod4, "\nSuggested threshold:", suggested\_threshold\_mod4, "\n")

## AUC for logmod4: 0.7732741   
## Suggested threshold: 0.4280508

# Compare the models  
if (auc\_mod3 > auc\_mod4) {  
 cat("Model logmod3 is better.\n")  
} else if (auc\_mod3 < auc\_mod4) {  
 cat("Model logmod4 is better.\n")  
} else {  
 cat("Both models are equal.\n")  
}

## Model logmod4 is better.

The model with a higher AUC is generally considered better for classification purposes. The suggested threshold is the one that maximizes sensitivity plus specificity on the ROC curve. It’s crucial to interpret the AUC along with the context of your specific problem and requirements.

1. (4pts) Compute the confusion matrix using logmod3 and logmod4 and a threshold of .25

confusion matrix using logmod3:

confusion matrix using logmod4:

Which model is better?

library(caret)  
  
# Predictions using logmod3 and logmod4 with threshold 0.25  
fit\_logmod3 <- predict(logmod3, newdata = newbreast, type = "response")  
fit\_logmod4 <- predict(logmod4, newdata = newbreast, type = "response")  
  
# Convert probabilities to binary predictions using threshold 0.25  
pred\_logmod3 <- ifelse(fit\_logmod3 > 0.25, 1, 0)  
pred\_logmod4 <- ifelse(fit\_logmod4 > 0.25, 1, 0)  
  
# True outcomes  
actual <- newbreast$Y  
  
# Confusion matrix for logmod3  
cm\_logmod3 <- confusionMatrix(data = as.factor(pred\_logmod3), reference = as.factor(actual))  
print("Confusion matrix using logmod3:")

## [1] "Confusion matrix using logmod3:"

print(cm\_logmod3$table)

## Reference  
## Prediction 0 1  
## 0 119 24  
## 1 77 57

# Confusion matrix for logmod4  
cm\_logmod4 <- confusionMatrix(data = as.factor(pred\_logmod4), reference = as.factor(actual))  
print("Confusion matrix using logmod4:")

## [1] "Confusion matrix using logmod4:"

print(cm\_logmod4$table)

## Reference  
## Prediction 0 1  
## 0 135 25  
## 1 61 56

To determine which model is better, you might consider metrics such as accuracy, precision, recall, or F1 score. In this case, accuracy is a commonly used metric: Accuracy for logmod3: (TP + TN) / (TP + TN + FP + FN) = (57 + 119) / (57 + 119 + 24 + 77) ≈ 0.648 Accuracy for logmod4: (TP + TN) / (TP + TN + FP + FN) = (56 + 135) / (56 + 135 + 25 + 61) ≈ 0.689 Considering accuracy as the evaluation metric, logmod4 appears to have a slightly higher accuracy compared to logmod3. Therefore, based on accuracy, logmod4 might be considered a better-performing model.

1. (2pt) Write a brief comment on the two models.

Both logmod3 and logmod4 are logistic regression models built to predict the binary outcome variable Y.

1. **logmod3:**
   * Utilizes menopause, tsize, node.caps, deg.malig, and irradiate as predictor variables.
   * The model includes linear terms for tsize and deg.malig.
   * Achieved an AUC of approximately 0.762.
2. **logmod4:**
   * Similar to logmod3 but introduces polynomial terms for tsize and deg.malig, including quadratic terms (tsize\_sq and deg\_malig\_sq).
   * Achieved a slightly higher AUC of approximately 0.773.

**Comparison:** - The introduction of polynomial terms in logmod4 allows the model to capture potential non-linear relationships between tsize, deg.malig, and the log-odds of the response variable. - Despite the slight improvement in AUC, the choice between logmod3 and logmod4 depends on factors such as interpretability, computational complexity, and the specific goals of the analysis. - Logmod4 may be preferred when there is evidence of non-linear relationships in the data, but it also comes with the cost of increased complexity. The decision should be made based on a balance between model performance and interpretability.