Homework 7 with SOLUTIONS

Directions. This homework pertains to materials on logistic regression in Lesson 6 and 7. The assignment should be typed, with your name on the document, and with properly labeled computer output. I suggest you attach your R/SAS input code at the end of your file clearly indicating the problem it corresponds to. If you choose to collaborate, the write-up should be your own. Please show your work! Upload the file to the HW7 Dropbox on ANGEL.

1. *50pts* Hastie and Tibshirani (1990) describe a study to determine risk factors for kyphosis, severe forward flexion of the spine following corrective spinal surgery. The age in months at the time of the operation for the 18 subjects for whom kyphosis was present were

   12 15 42 52 59 73 82 91 96 105 114 120 121 128 130 139 157

   and for 22 of the subjects for whom kyphosis was absent were

   1 1 2 8 11 18 22 31 37 61 72 81 97 112 118 127 131 140 151 159 177 206.

   (a) Fit a logistic regression model using age as a predictor of whether kyphosis is present. Test whether age has a significant effect.

   Solution: The logistic model fitted is

   $$\pi = \exp(\hat{\beta}_0 + \hat{\beta}_1 AGE)/[1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 AGE)$$

   or equivalently, $logit(\pi) = \hat{\beta}_0 + \hat{\beta}_1 AGE$.

   From the SAS/R output the estimate for $\hat{\beta}_1$ is 0.0043 with the p-value $0.463 > 0.05$. Therefore Age is not a significant covariate.

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.5727	0.6024	0.9038	0.3418
age	1	0.00430	0.00585	0.5393	0.4627

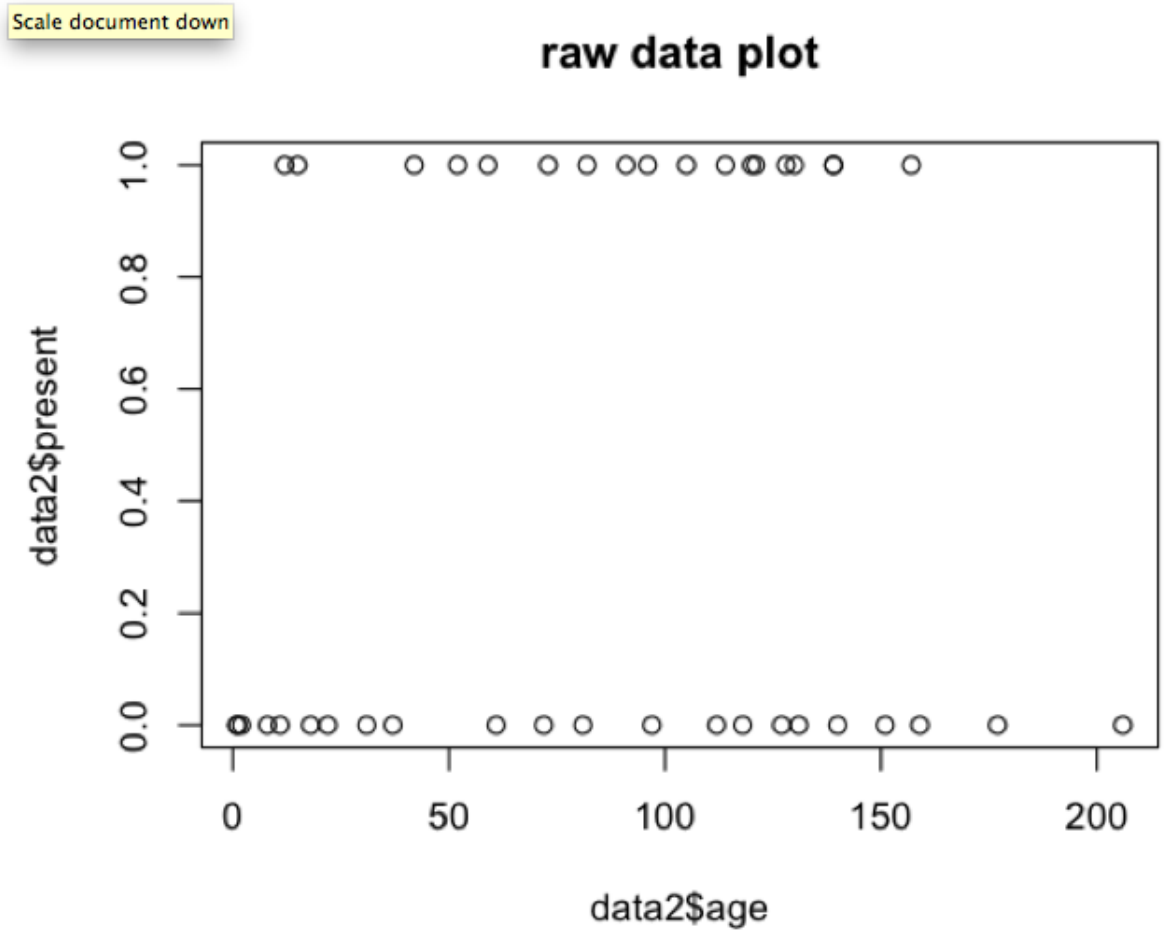
Odds Ratio Estimates

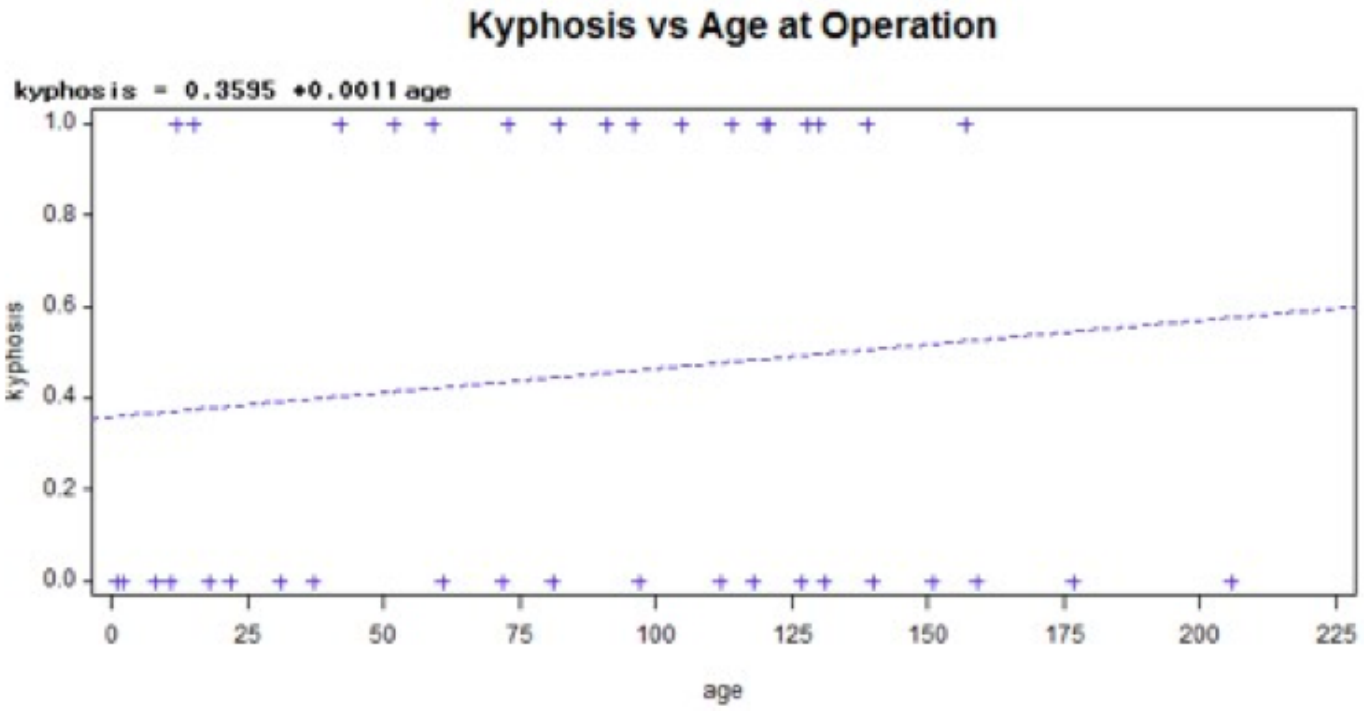
Effect	Point Estimate	95% Wald Confidence Limits	
age	1.004	0.993	1.016

(b) Plot the data. Note the difference in dispersion on age at the two levels of kyphosis. Do you need to adjust for dispersion? If yes, make the adjustment and report if there are any changes in your inference.

Solution:

The raw data plot is as follows. From the plot, we can see that the age has some distribution at the two levels of the kyphosis. Note that there is no overdispersion for ungrouped data, so we dont need to adjust. However, we can explore if there is some natural grouping based on the covariates that will allows to explore overdispersion.





If we look at the estimated scaling parameter, i.e., Pearson Chi-Square/ DF =1.1093. Thats also an indication of lack of overdispersion. We can aggregate data using **proc freq** in SAS or **table()** in R. We can also use option **AGGREGATE** in SAS, which calculates goodness-of-fit statistics for a table that aggregates over the unique patterns for the covariates appearing in the model by default, in this case just age. Here, I specify grouping by the variable age. From the output using option scale=pearson, we get the same scale parameter as above (i.e., there was no really additional grouping occuring) and we conclude that adjusting for overdispersion is not needed in this case.

(c) Fit the model $logit(\pi(x)) = \alpha + \beta_1 X + \beta_2 X^2$. Test the significance of the squared age term, plot the fit, and interpret.

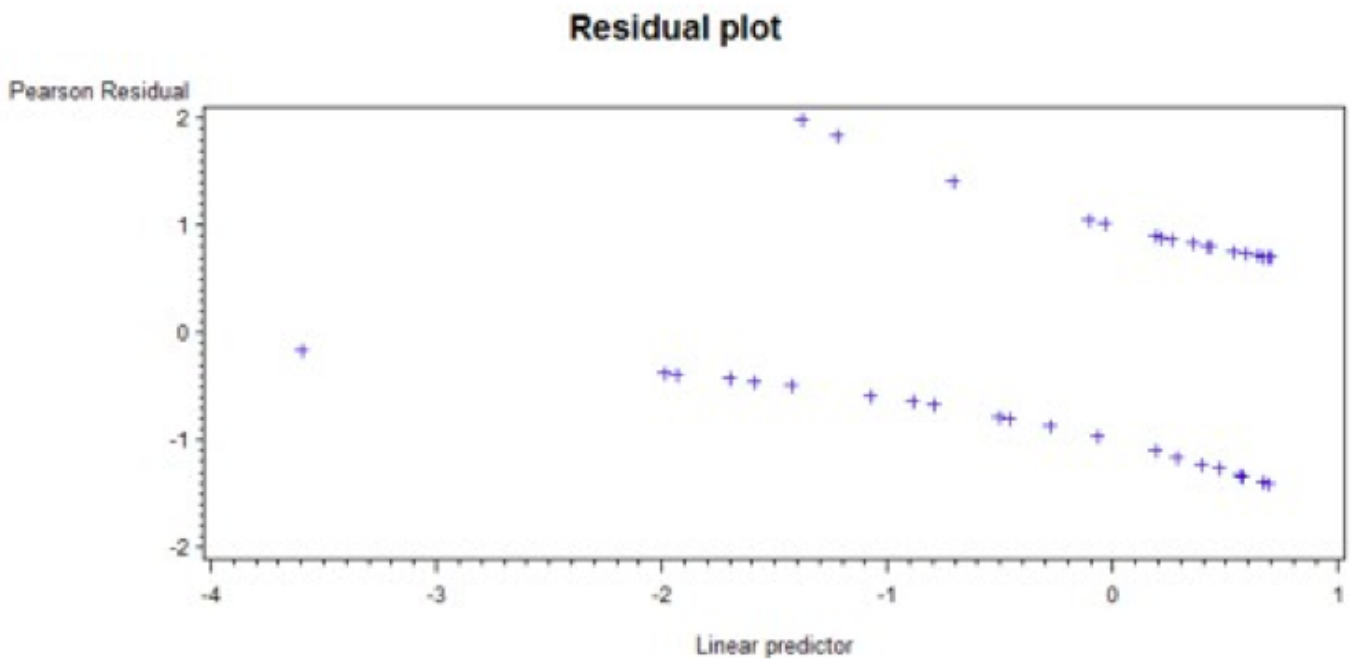
Solution: From the output the squared age term has a significant effect (p-value=0.036). The fit is plotted below. The residual plot is not very informative for ungrouped data. The $G^2 = 6.27$ with $df = 1$, which leads to p-value=0.012. So adding the squared age term improves the fit.

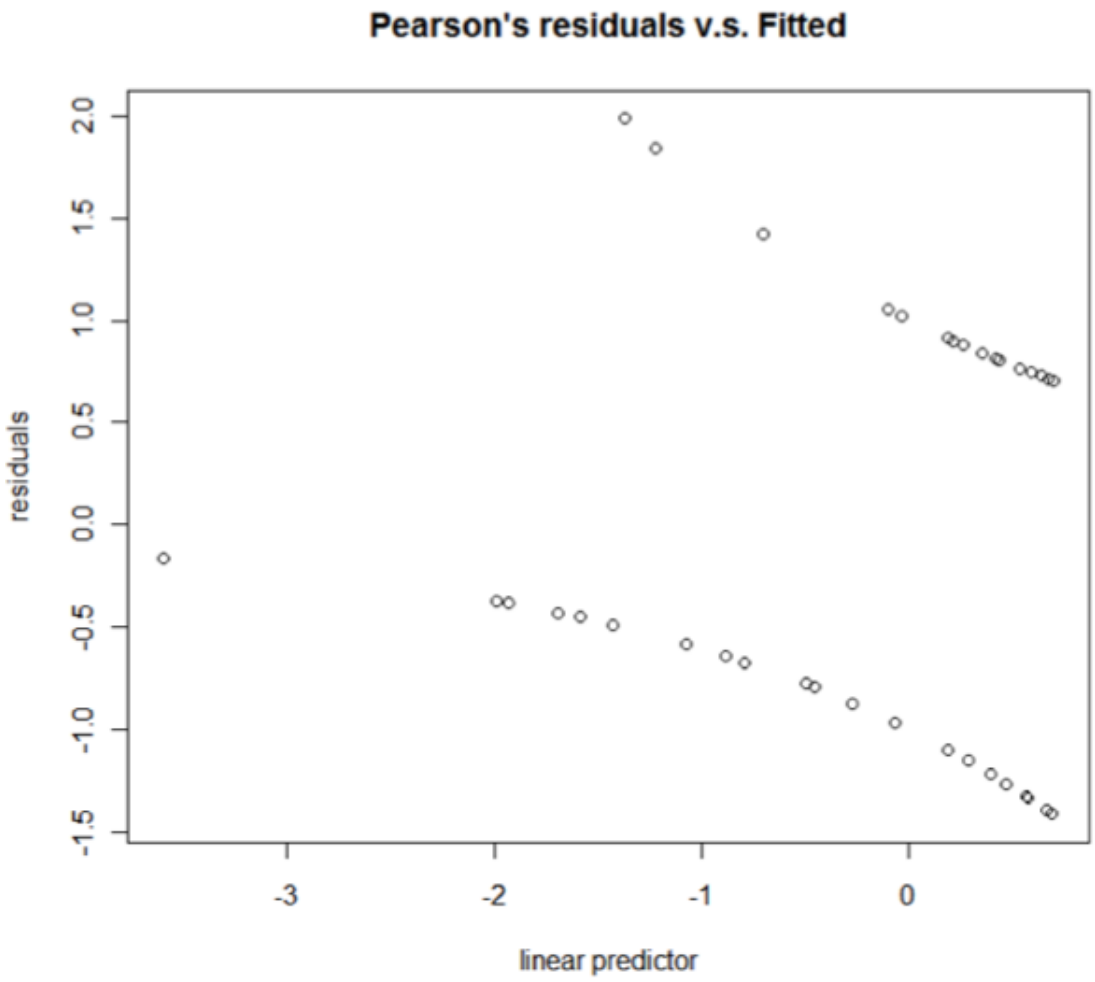
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.0463	0.9944	4.2348	0.0396
age	1	0.0600	0.0268	5.0259	0.0250
agesqr	1	-0.00033	0.000156	4.3960	0.0360

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age	1.062	1.008	1.119
agesqr	1.000	0.999	1.000





SAS code

```
data kyphosis;
infile "D:\stat504\UP\hw7\kyphosis.dat";
input age age2 kyphosis;
run;

proc reg data=kyphosis;
model kyphosis=age;
title 'Kyphosis vs Age at Operation';
plot kyphosis*age;
run;

proc logistic descending;
model kyphosis = age / lackfit;
```

```
run;

proc logistic descending;
model kyphosis = age / link=logit aggregate=(age) scale=pearson;
output out=out1 xbeta=xb reschi=reschi;
run;

axis1 label=('Linear predictor');
axis2 label=('Pearson Residual');
proc gplot data=out1;
title 'Residual plot';
plot reschi * xb / haxis=axis1 vaxis=axis2;
run;


proc logistic descending;
model kyphosis = age age2/ link=logit aggregate scale=pearson;
output out=out2 xbeta=xb2 reschi=reschi2;
run;

axis1 label=('Linear predictor');
axis2 label=('Pearson Residual');
proc gplot data=out2;
title 'Residual plot';
plot reschi2 * xb2 / haxis=axis1 vaxis=axis2;
run;
```

R Code/Output:

```
##(a)
> names(data2)=list("age","age2","presence")
> res1=glm(presence~age,family=binomial(link=logit),data=data2)
> summary(res1)



Call:

glm(formula = kyphosis ~ age,
```

```
family = binomial(link = "logit"))

Deviance Residuals:

Min 1Q Median 3Q Max

-1.3126 -1.0907 -0.9482 1.2170 1.4052

Coefficients:

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

(Intercept) -0.572693 0.602395 -0.951 0.342

age 0.004296 0.005849 0.734 0.463

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 55.051 on 39 degrees of freedom

Residual deviance: 54.504 on 38 degrees of freedom

AIC: 58.504

Number of Fisher Scoring iterations: 4


## (b)
> 1-pchisq(res1$dev,38)##(c)
> res2=glm(presence~age+age2,family=binomial(link=logit),data=data2)
> plot(res2$linear.predictors, residuals(res2, type="pearson"),
main="Pearson's residuals v.s. Fitted")


Call:
glm(formula = kyphosis ~ age + agesq,
family = binomial(link = "logit"))
Deviance Residuals:

Min 1Q Median 3Q Max

-1.482 -1.009 -0.507 1.012 1.788
```

```
Coefficients:

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

(Intercept) -2.0462547 0.9943478 -2.058 0.0396 *
age 0.0600398 0.0267808 2.242 0.0250 *
agesq -0.0003279 0.0001564 -2.097 0.0360 *
---
Signif. codes: 0 ?***?0.001 ?**?0.01 ??0.05 ??0.1 ??1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 55.051 on 39 degrees of freedom
Residual deviance: 48.228 on 37 degrees of freedom
AIC: 54.228
Number of Fisher Scoring iterations: 4
```

2. *50pts* Some additional diagnostic tools for logistic regression are described in Lesson 6 and also in Agresti(2007) 5.1.6-5.1.8, and Agresti(2013) Section 6.3. A *receiver operating characteristic* (ROC) curve and its associated statistics is one of the way to assess predictive power of your logistic regression model. You can refer to the Accuracy-Handout.pdf on ANGEL as well, or consult other sources. For SAS, you can use option `OUTROC=` in `MODEL` specification of `PROC LOGISTIC`.

   Then you need to use either `GPLOT` or `ODS` Graphics to plot data from `roc1`.

   For exxample

```
ods html;
  ods graphics on;

  proc logistic descending data=donner;
   model survive = age / lackfit influence iplots outroc=roc1;
 /*      units age=10;*/
  run;

  ods graphics off;
  ods html close;
```

   For more details see:

   http://support.sas.com/onlinedoc/913/getDoc/enstatug.hlp/ logistic_sect37.htm#stat_logistic_logisticrocc

http://support.sas.com/onlinedoc/913/getDoc/en/statug.hlp/logistic_sect60.htm#stat_logistic_logisticex8

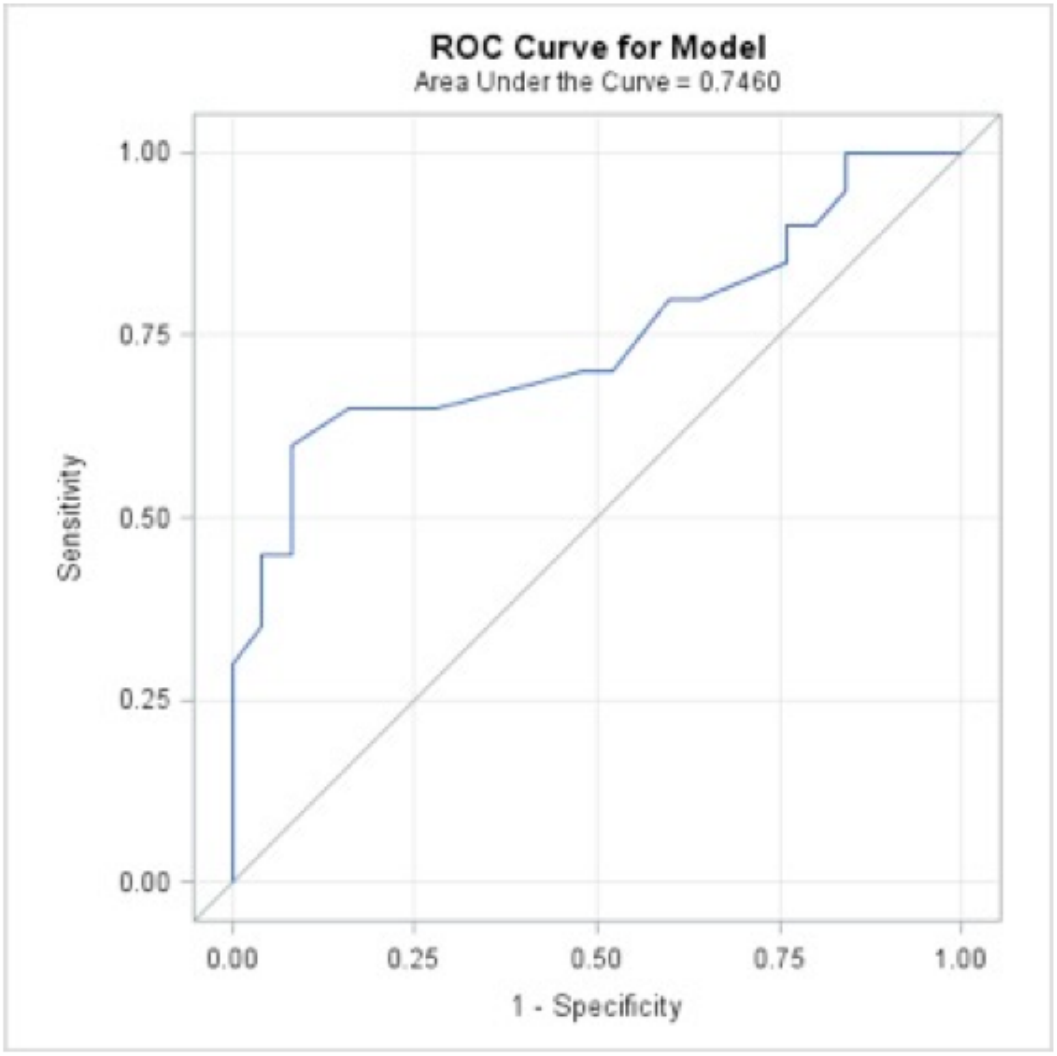
For R, use ROC & Hosmer-Lemeshow.R file on ANGEL.

For logistic regression you can create a 2×2 classification table of predicted values from your model for your response if $\hat{y} = 0$ or 1 versus what the true value of $y = 0$ or 1. The prediction if $\hat{y} = 1$ depends on some cut-off probability, π_0. For example, $\hat{y} = 1$ if $\hat{\pi}_i > \pi_0$ and $\hat{y} = 0$ if $\hat{\pi}_i \leq \pi_0$. The most common cut-off value is $\pi_0 = 0.5$. Then $sensitivity = P(\hat{y} = 1|y = 1)$ and $specificity = P(\hat{y} = 0|y = 0)$.

(a) Fit the ROC curve for the DONNER data (donner.dat or donner.txt on ANGEL) for the main effects logistic regression model. Turn in the plot and report the A value. What can you say about the predictive power of your model?

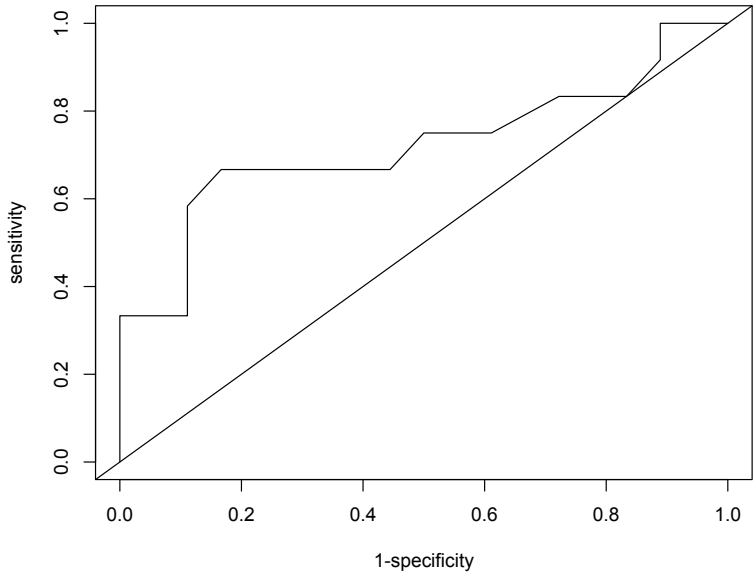
Solution:

The ROC curve is shown below. From the above output, the A value (i.e., Area Under the Curve) is 0.746, indicating that the model has good predictive power.



(b) Now you will do the simplest case of model validation known as *cross-validation*. First, randomly select 30 out of the 45 cases from donner.txt, and save this in a new dataset labeled as donnerTrain.txt. Save the remaining 10 observations in a dateset labeled donnerTest.txt. Fit the main effects logistic regression model based on donnerTrain.txt. Plot the ROC curve and report the A value.

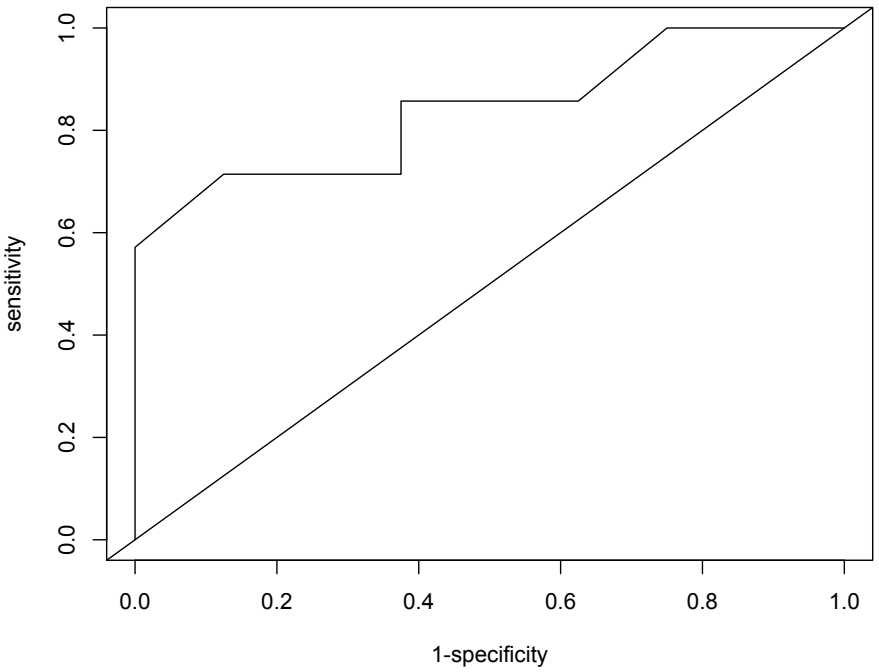
Solution: REMARK: for part (c) and (d), your result can be different due to the sampling noise.
The ROC curve is shown below. From the above output, the A value (i.e., Area Under the Curve) is 0.7199, indicating that the model has good predictive power.



(c) Use the fitted model from part (b) to predict the probability of survival for observations in donnerTest.txt. Fit the ROC curve and report the A-value for this test dataset. How does the predictive power of your model compare between the training and test datasets.

Solution:

Predict the probability of survival for observations in testing dataset. The A-value for test dataset is 0.84. The predictive power for training data is higher than for test data. This is expected, because the model is trained on training data.



R Code:

```
> donner=read.table("donner.txt")
> survive=donner[,3]
> age=donner[,1]
> sex=donner[,2]
>
> result=glm(survive~age+sex,family=binomial("logit"))
>
> roc.plot <-
+ function (sd, sdc, newplot = TRUE, ...)
+ {
+     sall <- sort(c(sd, sdc))
+     sens <- 0
+     specc <- 0
+     for (i in length(sall):1) {
+         sens <- c(sens, mean(sd >= sall[i], na.rm = T))
+         specc <- c(specc, mean(sdc >= sall[i], na.rm = T))
+     }
+     if (newplot) {
+  plot(specc, sens, xlim = c(0, 1), ylim = c(0, 1), type = "l",
+  xlab = "1-specificity", ylab = "sensitivity", ...)
```

```
+  abline(0, 1)
+     }
+     else lines(specc, sens, ...)
+     npoints <- length(sens)
+  area <- sum(0.5 * (sens[-1] + sens[-npoints]) * (specc[-1] -
 specc[-npoints]))
+  lift <- (sens - specc)[-1]
+     cutoff <- sall[lift == max(lift)][1]
+     sensopt <- sens[-1][lift == max(lift)][1]
+     specopt <- 1 - specc[-1][lift == max(lift)][1]
+  list(area = area, cutoff = cutoff, sensopt = sensopt,
 specopt = specopt)}
> roc.analysis <-
+ function (object, newdata = NULL, newplot = TRUE, ...)
+ {
+     if (is.null(newdata)) {
+  sd <- object$fitted[object$y == 1]
+  sdc <- object$fitted[object$y == 0]
+     }
+ else {
+ sd <- predict(object, newdata, type = "response")[newdata$y ==
+ 1]
+ sdc <- predict(object, newdata, type = "response")[newdata$y ==
+ 0]
+     }
+  roc.plot(sd, sdc, newplot, ...)
+ }
>
> roc.analysis(result)
$area
[1] 0.746

$cutoff
       32
0.3363082

$sensopt
[1] 0.6

$specopt
[1] 0.92

> x=1:45
> index=sample(x,30,replace=F)
> donnerTrain=donner[index,]
```

```
> colnames(donnerTrain)=c("age","sex","survive")
> donnerTrain=as.data.frame(donnerTrain)
> donnerTest=donner[-index,]
> colnames(donnerTest)=c("age","sex","survive")
> donnerTest=as.data.frame(donnerTest)
>
> ##fit the training model
> result.train=glm(survive~age+sex, family=binomial("logit"),
data=donnerTrain)
> roc.analysis(result.train)
$area
[1] 0.658371

$cutoff
       30
0.3823259

$sensopt
[1] 0.5384615

$specopt
[1] 0.8235294

> donnerTest=read.table('donnerTest.txt')
> #c
> predict.test=predict(result.train, donnerTest, type='response')
> print(predict.test)
        1         3         6         9        11        13
 16        18        21        24
0.4671048 0.2735448 0.2735448 0.0980333
0.6946247 0.4060587 0.7152937 0.7889910 0.7446620 0.1223074
       25        31        33        37        45
0.4424621 0.5660733 0.7350972 0.3823259 0.6946247
> result.test = glm(donnerTest[,3]~donnerTest[,1]+
donnerTest[,2],family=binomial("logit"))
> roc.analysis(result.test)
$area
[1] 0.8392857

$cutoff
       14
0.2447245

$sensopt
[1] 0.7142857
```

```
$specopt
[1] 0.875

>
```

SAS Code:

```
proc survey
select data=donner out=split samprate=.333 outall;
run;

data random1 random2;
set split;
if selected = 1 then
output random1;else output random2;run;

ods html;
ods graphics on;
proc logist descending data=donner_train;
class Survive / order=data param=ref ref=first;
model Survive = Age Gender / lackfit influence iplots outroc=roc1;
score data=donner_test outroc=roc_score;
run;
ods graphics off;
ods html close;
```