# Department of Statistics

## COURSE STATS 330/762

Model answers Assignment 5, 2010

Instructions: Hand in your completed assignment to the Student Resource Centre by **4pm October 14th.**

The data set for this assignment is in the file **acath.txt** which is available on the course web page.

1. *Read the data into R, and make a data frame. Check for gross errors. (Note that, due to the size of this data set, I have not reproduced it at the end of the assignment. I have not deliberately introduced any errors into the data. ) Print out the first 16 lines. [5 marks]*

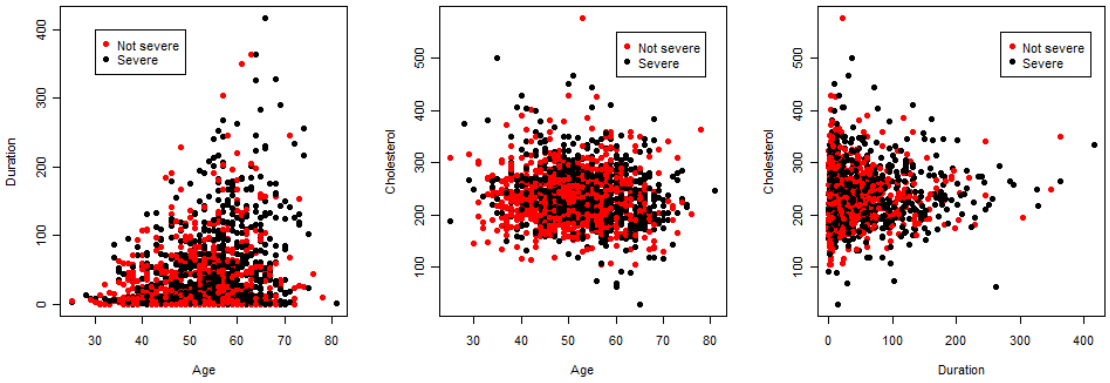
A pairs plot shows no obvious outliers. The usual code will read in the data and print out the first 16 lines. [5 marks: 2 for reading in, 2 for checks, 1 for printing]

2. *Perform a graphical analysis of the data, without fitting any model, that will let you see how the risk factors sex, age, cad.dur and choleste affect the probability of having severe coronary artery disease . [10 marks]*

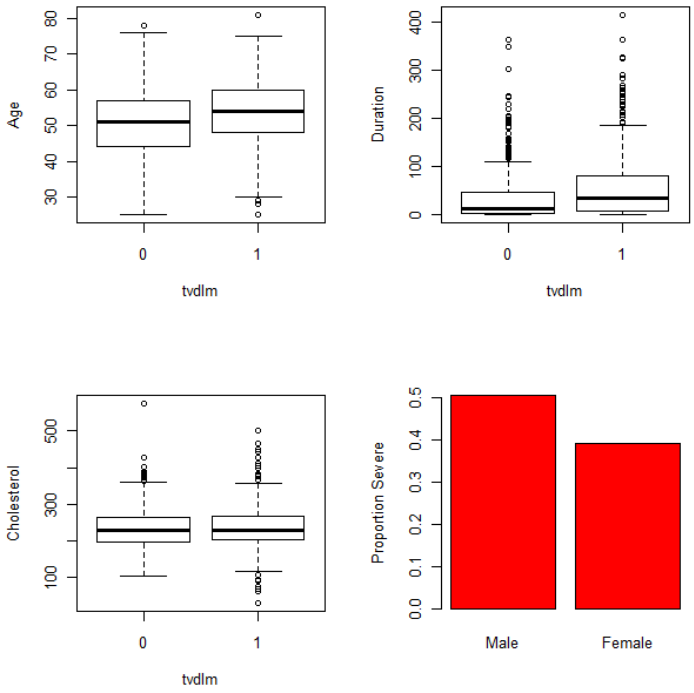
**Drawing some plots:** Below we show some exploratory plots of these data. These were produced by the code

```
par(mfrow=c(1,3))
my.pch=c(19, 22)
my.col = c("red","black")
plot(acath.df$age, acath.df$cad.dur, type="n", xlab="Age", ylab =
"Duration")
points(acath.df$age, acath.df$cad.dur, pch=19, col=
my.col[acath.df$tvdlm+1])
legend(30,400, legend=, c("Not severe","Severe"), pch=19, col=my.col)

plot(acath.df$age, acath.df$choleste, type="n", xlab="Age", ylab =
"Cholesterol")
points(acath.df$age, acath.df$choleste, pch=19, col=
my.col[acath.df$tvdlm+1])
legend(60,550, legend=, c("Not severe","Severe"), pch=19, col=my.col)

plot(acath.df$cad.dur, acath.df$choleste, type="n", xlab="Duration",
ylab = "Cholesterol")
points(acath.df$cad.dur, acath.df$choleste, pch=19, col=
my.col[acath.df$tvdlm+1])
legend(270,550, legend=, c("Not severe","Severe"), pch=19, col=my.col)
```

```
par(mfrow=c(2,2))
boxplot(age~tvdlm,  xlab="tvdlm", ylab = "Age", data = acath.df)
boxplot(cad.dur~tvdlm,  xlab="tvdlm", ylab = "Duration", data =
acath.df)
boxplot(choleste~tvdlm,  xlab="tvdlm", ylab = "Cholesterol", data =
acath.df)
props = tapply(acath.df$tvdlm, acath.df$sex, mean)

barplot(props, names.arg=c("Male","Female"), col="red", ylab =
"Proportion Severe")
```

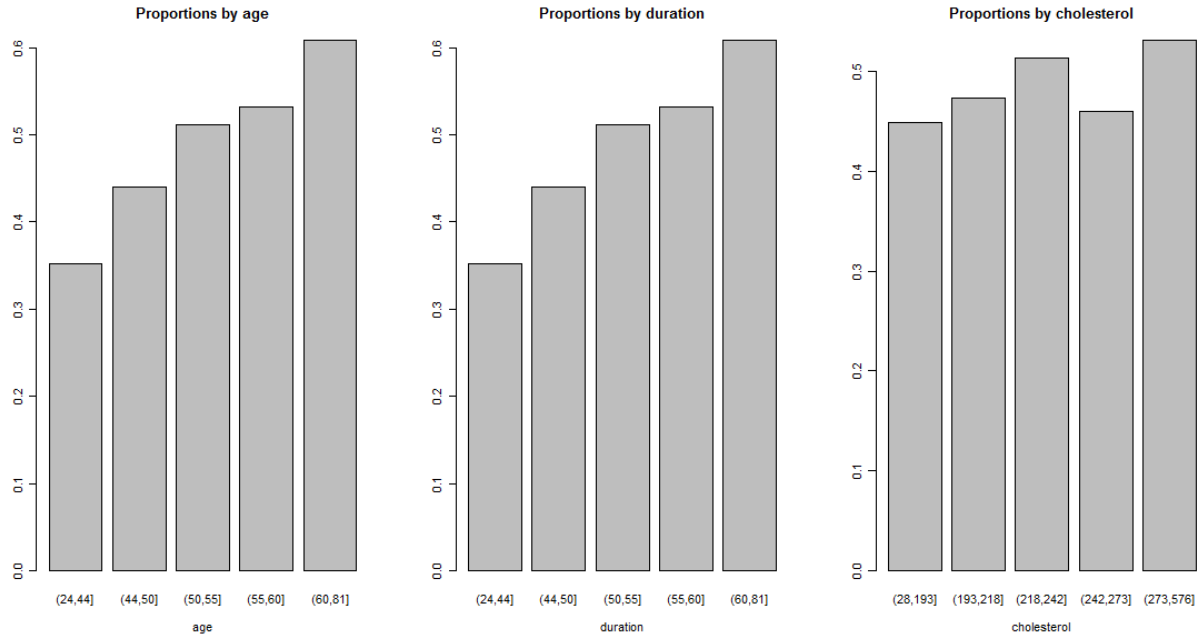



Seems that there are more black dots than red for higher values of Age and Duration.



We can also convert the continuous variables into groups and plot them as we did in Assignment 1:

```
#   barplots of age etc groups

par(mfrow=c(1,3))
cut.points = quantile(acath.df$age, prob = seq(0,1,by=0.2))
cut.points[1]=cut.points[1]-1
age.group = cut(acath.df$age, cut.points )
prob.severe = tapply(acath.df$tvdlm, age.group,mean)
barplot(prob.severe, main = "Proportions by age", xlab="age")

cut.points = quantile(acath.df$cad.dur, prob = seq(0,1,by=0.2))
cut.points[1]=cut.points[1]-1
cad.group = cut(acath.df$cad.dur, cut.points )
prob.severe = tapply(acath.df$tvdlm, cad.dur.group,mean)
barplot(prob.severe, main = "Proportions by duration", xlab =
"duration")

cut.points = quantile(acath.df$choleste, prob = seq(0,1,by=0.2))
cut.points[1]=cut.points[1]-1
chol.group = cut(acath.df$choleste, cut.points )
prob.severe = tapply(acath.df$tvdlm, chol.group,mean)
barplot(prob.severe, main = "Proportions by cholesterol", xlab =
"cholesterol")
```




[5/10 for suitable plots]

**Interpretation:** Based on these plots, it seems that the variables Sex, Age and Duration have a strong effect on the chance of severe disease, but the effect of Cholesterol is a bit weaker. Males are more likely to have a serious condition, as are older patients and patients with longer duration of symptoms. [5/10 for interpretation]

3. *Fit a logistic regression model to the data, diagnosing any major problems with the fit. Interpret the coefficients. Does your interpretation confirm the conclusions you reached in question 2? [15 marks]*

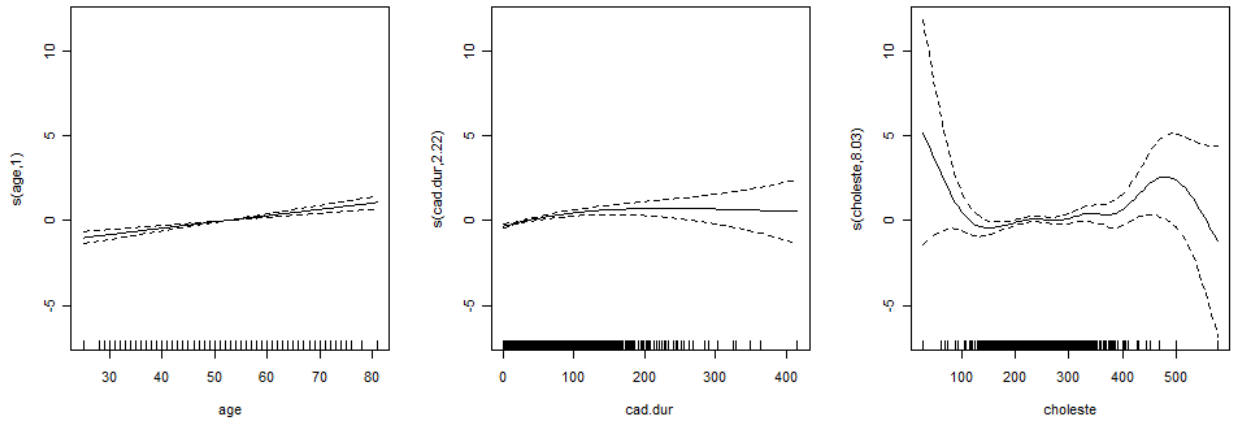
**Fitting the model:** Fitting the basic model, with interactions between the continuous variable and sex gives gives

```
> acath.glm = glm(tvdlm ~ age*sex + cad.dur*sex + choleste*sex,
family=binomial, data=acath.df)
> summary(acath.glm)

Call:
glm(formula = tvdlm ~ age * sex + cad.dur * sex + choleste *
    sex, family = binomial, data = acath.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1146  -1.0716  -0.7661   1.1660   1.9667

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.030672   0.523068  -5.794 6.87e-09 ***
age           0.036594   0.007365   4.969 6.75e-07 ***
sex          -0.141424   1.083548  -0.131 0.896155
cad.dur       0.007773   0.001318   5.895 3.74e-09 ***
choleste      0.003614   0.001264   2.859 0.004248 **
age:sex       0.006938   0.016296   0.426 0.670305
sex:cad.dur  -0.009686   0.002804  -3.454 0.000553 ***
sex:choleste -0.001945   0.002297  -0.847 0.397110
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2064.3  on 1489  degrees of freedom
Residual deviance: 1945.7  on 1482  degrees of freedom
AIC: 1961.7

Number of Fisher Scoring iterations: 4
```

Thus, all variables seem significant, even choleste. Note sex is scored 1=female, 0= male so that the negative sign is consistent with our exploratory graphs. Interpreting the coefficients confirms the impression gained from our exploratory plots. There is also a significant interaction between sex and duration. The HL statistic is not significant:

```
> HLstat(acath.glm)
Value of HL statistic =  11.533
P-value =  0.173
```

Let's try a gam plot to see if any transformations need be made:




Looks like age is OK, what about duration and choleste? We could fit a polynomial in duration, but the interactions complicate things a bit. An alternative is to log the durations. A gam plot of this variable indicates no further transformations of durations are required. We could try a cubic in choleste, but the 2 and 3rd degree terms are not significant. Accordingly we stick with the model `tvdlm ~ age +sex*lcd + choleste:`

```
acath4.glm = glm(tvdlm ~ age +sex*lcd +  choleste, family=binomial,
data=acath.df)
```

The summary is

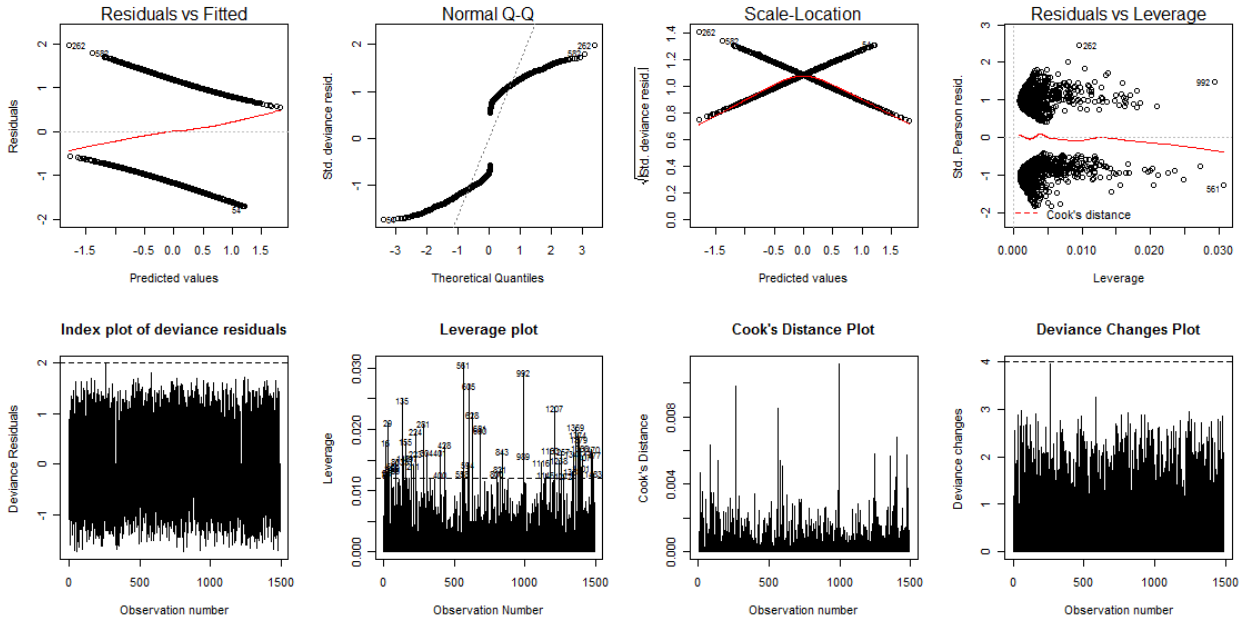
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.532033   0.450438  -7.841 4.46e-15 ***
age          0.037411   0.006493   5.761 8.34e-09 ***
sex          0.208614   0.363462   0.574  0.56599
lcd          0.310878   0.044958   6.915 4.68e-12 ***
choleste     0.002969   0.001056   2.811  0.00494 **
sex:lcd     -0.287069   0.108520  -2.645  0.00816 **
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2064.3  on 1489  degrees of freedom
Residual deviance: 1937.5  on 1484  degrees of freedom
AIC: 1949.5
```

An alternative model, favoured by several students, was to omit the interactions, but include a quadratic term in cad.dur. This had an AIC of 1963.8, so was not quite as good as the model above.

[5/15 marks for fitting the model. Deduct 2 if just the basic additive model is fitted.]

**Diagnostic plots:** Let's do some diagnostic plots:

```
> par(mfrow=c(2,4))
> plot(acath4.glm)
> glm.diag.plots(acath3.glm)
```




Looks like points 561 and 992 might cause problems. Let's look at the effect on the coefficients, comparing the model with all the data and with these 2 points removed:

```
> acath5.glm = glm(tvdlm ~  age +sex*lcd +  choleste,
+ subset = -c(561,992), family=binomial, data=acath.df)
> cbind(coef(acath4.glm), coef(acath5.glm))
```

```
            All points in  561 & 992 out
(Intercept) -3.532032756   -3.607796980
age          0.037411218    0.037402441
sex          0.208614299    0.136957317
lcd          0.310878316    0.310803761
choleste     0.002969143    0.003300499
sex:lcd     -0.287068833   -0.267081869
```

The effect is not too bad. We will look at the effect on the curves in the next question. [5/15 marks for diagnostics]

**Interpretation:** To interpret the effect of age, and cholesterol, we note that the coefficients are positive and so the log-odds (and hencethe odds and the probabilities) of severe disease both go up as age and cholesterol increase. The interaction between sex and lcd complicates the interpretation. For the males (ie baseline), the log odds goes up by 0.310878316 for each unit increase in lcd. For the females, the log odds goes up by

0.310878316 -0.287068833 = 0.02380948 for each unit increase in lcd (i.e. hardly at all). [5/15 marks for interpretation]

4. *Produce a graph that will illustrate how the fitted probabilities for your model change with age and sex. (You can fix the values of cad.dur and choleste at their averages.) Is the risk higher for males? [10 marks]*

**Plotting:** The age range is from 25 to 81. Our strategy is to use the predict function to calculate the fitted probabilities for each age/sex combination, then draw the picture. The first step is to calculate a suitable data frame containing the age/sex values. We will fix the other variables at their means. The following code creates the required data frame:

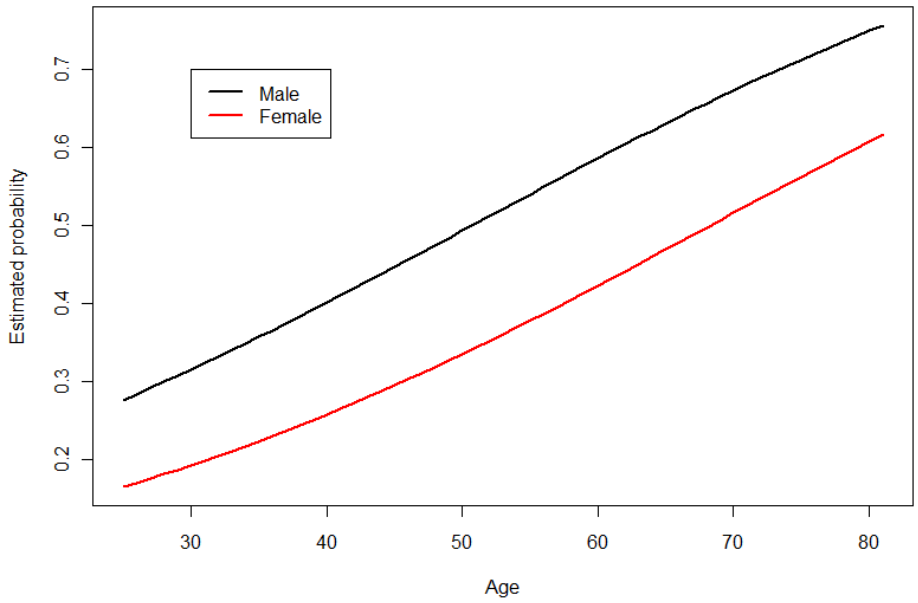
```
age.points = 25:81
newdata = data.frame( sex = rep(c(0,1), length(age.points)),
age = rep(age.points, each=2),
cad.dur = mean(lcd), choleste=mean(acath.df$choleste))
```

The first 10 lines (out of 114 in total) are

```
   sex age      lcd choleste
1    0  25 3.018427 234.9866
2    1  25 3.018427 234.9866
3    0  26 3.018427 234.9866
4    1  26 3.018427 234.9866
5    0  27 3.018427 234.9866
6    1  27 3.018427 234.9866
7    0  28 3.018427 234.9866
8    1  28 3.018427 234.9866
9    0  29 3.018427 234.9866
10   1  29 3.018427 234.9866
```

The following code draws the picture, using both the full data and the data with 2 points removed:

```
> pred = predict(acath4.glm, newdata=newdata, type="response")
> par(mfrow=c(1,1))
> plot(newdata$age , pred, type="n", xlab="Age",
      ylab = "Estimated probability")
> lines(newdata$age[newdata$sex==0], pred[newdata$sex==0],
      col="black", lwd=2)
> lines(newdata$age[newdata$sex==1], pred[newdata$sex==1],
     col="red", lwd=2)
```




[5/10 for the plot. A trellis-style plot is OK, as long as the scales are the same]

Interpretation: The graphs drawn with the two data points removed are virtually identical, so we leave the points in. Looks like males with average log duration and cholesterol have a higher risk than females. [5/10 for interpretation]

5. *Extra question for 762 only: Do you think the model you have chosen has any value as a predictive tool? [10 marks]*

Calculate sensitivity and specificity: Calculate the sensitivity and specificity using cross validation:

```
> cross.val.glm(formula(acath4.glm),
   data=data.frame(acath.df, lcd), nfold=50)
Mean Specificity =  0.6708793
Mean Sensitivity =  0.5626054
Mean Correctly classified =  0.6183448
```

[5/10 marks for the calculation]

**Interpretation:** Thus the model correctly classifies 67% of the non-severe cases and 56% of the severe. Not a great classifier, particularly for the severe cases, only a bit better than random guessing. [5/10 marks for interpretation]

The sensitivity and specificity worked out on the data (not using cross-validation) are obtained using the code

```
> my.pred =   predict(acath4.glm) >0  #  TRUE means predicting a
"success", FALSE a failure


> table(acath.df$tvdlm, my.pred)
   my.pred
    FALSE TRUE
  0   519  248
  1   314  409
```

Thus the specificity is 519/(519+248) = 0.6766623 and the sensitivity is 409/(314+409) = 0.5656985. These values are pretty close to the cross-validated ones.

[Deduct 2 marks if no cross-validation is done.]