Generalized Linear Models

Dr. Prabhanjan Tattar
Lead Statistician
CustomerXPs Software Private Limited
Bangalore

The Topics

- Logistic Regression
- Probit Regression
- Poisson Regression (Log-linear Regression)

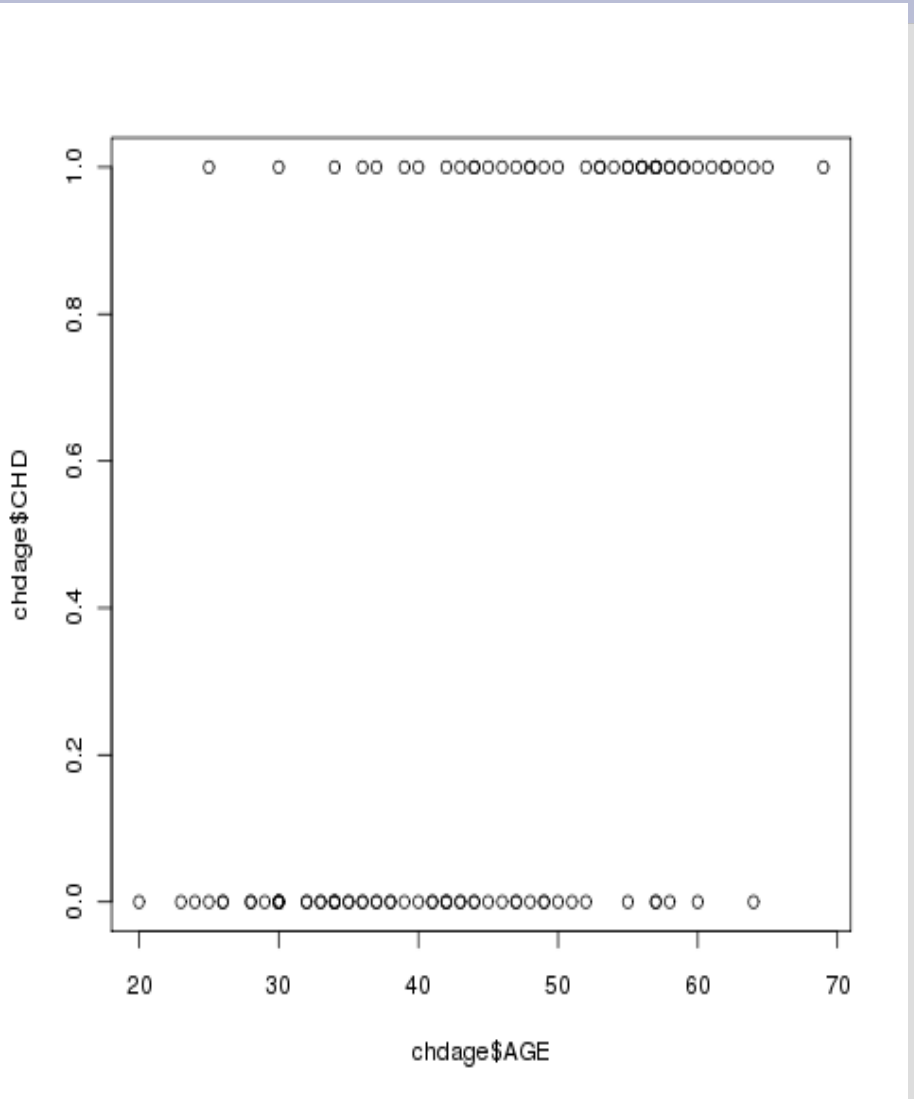
The Classics

- Hosmer and Lemeshow (2000)
- Kleinbaum and Klein (200x)
- Dobson (2002)
- McCullagh and Nelder (1989)

Coronary Heart Disease

- Age of Patients and Coronary Heart Disease
- 100 subjects in the study
- chdage$CHD taking value 1 indicates presence of the disease, and 0 the absence of it

```
tiff("CHD_AGE.tiff")
plot(chdage$AGE,chdage$CHD)
dev.off()
null device
```



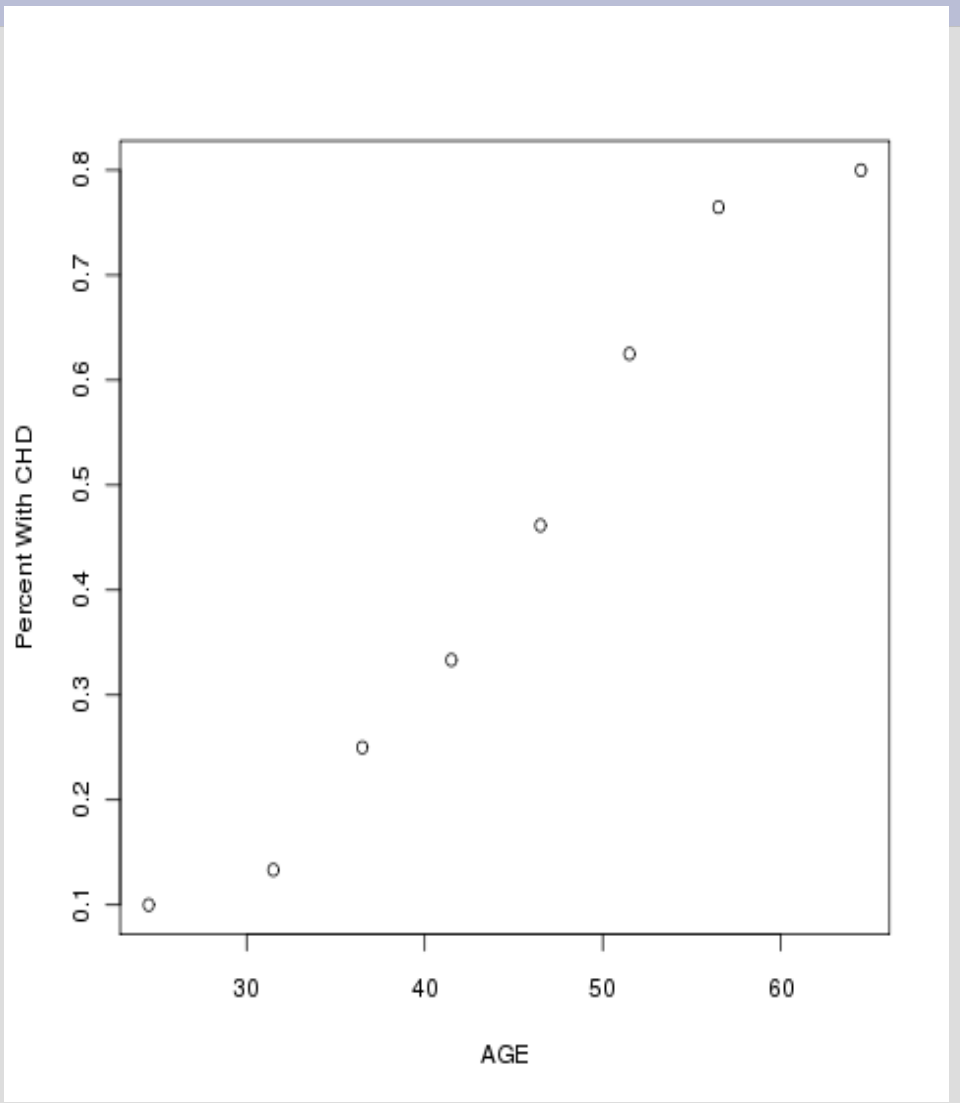
- The interest is in understanding the relationship between the age and presence or absence of CHD
- Scatterplot shows two parallel lines with intercepts at 0 and 1 respectively
- Note that the variability in CHD at all the age points is large

```
agegrp=cut(chdage$AGE,c(19,29,
34,39,44,49,54,59,69),include.lo
west=T,labels=c("20-29","30-34",
"35-39","40-44","45-49","50-54",
"54-59","60-69"))
table(agegrp,chdage$CHD)
```

agegrp	0	1
20-29	9	1
30-34	13	2
35-39	9	3
40-44	10	5
45-49	7	6
50-54	3	5
54-59	4	13
60-69	2	8

```
xx=cbind(x,rowSums(x[,1:2]),x[,2]/rowSums(x[,1:2]))
colnames(xx)=c("CHD=0","CHD=1","n","Proportion")
xx
```

	CHD=0	CHD=1	n	Proportion
20-29	9	1	10	0.1
30-34	13	2	15	0.13
35-39	9	3	12	0.25
40-44	10	5	15	0.33
45-49	7	6	13	0.46
50-54	3	5	8	0.63
54-59	4	13	17	0.76
60-69	2	8	10	0.8



The Logistic Regression Model

- For sure, the previous graph exhibits linearity
- Assures that $E(Y|x)$ will be between 0 and 1
- As a variable of the age, the curve promises to reach 0 or 1 gradually and eventually
- The "curve" also looks *S*-shaped
- With many theoretical possibilities for such curve, the popular choice has been *logistic curve*

- The *logistic regression* model is

$$\pi(x)=\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$

- Under this model, Y_i takes the value 1 or 0 with probabilities $\pi(x)$ and $1-\pi(x)$ respectively

That is,

$$P(Y=1|x)=\pi(x)=\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$

$$P(Y=0|x)=\pi(x)=\frac{1}{1+e^{(\beta_0+\beta_1 x)}}$$

- A useful transformation is

$$g(x)=\ln\left[\frac{\pi(x)}{1-\pi(x)}\right]$$
$$=\beta_0+\beta_1 x$$

- Of course, *g(x)* is also known as the *logit* transformation

- The model here is $y=\pi(x)+\epsilon$, with the error taking only two values 0 and 1
- If $y = 1$, $\epsilon=1-\pi(x)$ with probability $\pi(x)$
- If $y = 0$, $\epsilon=-\pi(x)$ with probability $1-\pi(x)$
- Thus, ϵ has a binomial distribution with mean zero and variance

$$\pi(x)(1-\pi(x))$$

The Likelihood Function

The contribution of the *i*-th datum would be

$$\pi(x_i)^{y_i} \times [1-\pi(x_i)]^{1-y_i}$$

and thus, the likelihood based on *n*- observations will be

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \times [1-\pi(x_i)]^{1-y_i}$$

The log-likelihood function is thus

$$\ln(l(\beta)) = \sum_{i=1}^{n} y_i \ln[\pi(x_i)] + \sum_{i=1}^{n} (1-y_i) \ln[1-\pi(x_i)]$$

The *normal equations* are then given by

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0$$

$$\sum_{i=1}^{n} x_i [y_i - \pi(x_i)] = 0$$

CHD Example Revisited

- Fitting the model using R, we get

```
chdglm=glm(chdage$CHD~chdage$AGE,family='binomial')
chdglm$coefficients
(Intercept)  chdage$AGE
 -5.309    0.111
```

That is, the fitted model is then

$$\pi(x)=\frac{e^{(-5.309+0.111x)}}{1+e^{(-5.309+0.111x)}}$$

One way of checking the fit of the model is verifying the equation

$$\sum_{i=1}^{n} y_i=\sum_{i=1}^{n} \hat{\pi}(x_i)$$

```
pix = function(x) {exp(-5.309 + 0.111*x)/(1+exp(-5.309 + 0.111*x))}
sum(pix(chdage$AGE))
[1] 43.07326
 sum(chdage$CHD)
[1] 43
```

We can use binomial(logit) instead of family='binomial'

CHD Example: LR Model Fitted

```
summary(chdglm)
Call:
glm(formula = chdage$CHD ~ chdage$AGE, family = "binomial")
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9718  -0.8456  -0.4576   0.8253   2.2859
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.30945    1.13365  -4.683 2.82e-06 ***
chdage$AGE   0.11092    0.02406   4.610 4.02e-06 ***
---
Signif. codes:  0 â***â 0.001 â**â 0.01 â*â 0.05 â.â 0.1 â â 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.35  on 98  degrees of freedom
AIC: 111.35
Number of Fisher Scoring iterations: 4
```

- The summary of the logistic regression model gives us most of what we want

CHD Example: Confidence Intervals

- Confidence intervals are missing though
- 100(1-a)% C.I for the slope and intercept terms (of the logit function) are given by

$$\hat{\beta}_1 \pm z_{1-\alpha/2} \hat{SE}(\hat{\beta}_1)$$

$$\hat{\beta}_0 \pm z_{1-\alpha/2} \hat{SE}(\hat{\beta}_0)$$

```
ucl_intercept=schdglm$coefficients[,1][1] + qnorm(1-0.05/2)*schdglm$coefficients[,2][1]
(Intercept)
  -3.087533
lcl_intercept=schdglm$coefficients[,1][1] - qnorm(1-0.05/2)*schdglm$coefficients[,2][1]
(Intercept)
  -7.531374
ucl_slope=schdglm$coefficients[,1][2] + qnorm(1-0.05/2)*schdglm$coefficients[,2][2]
chdage$AGE
 0.1580775
lcl_slope=schdglm$coefficients[,1][2] - qnorm(1-0.05/2)*schdglm$coefficients[,2][2]
chdage$AGE
0.06376477
```

CHD Example: CI for Logit Function

- The estimate of the logit function is given by

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

and its variance by

$$\hat{Var}[\hat{g}(x)] = \hat{Var}(\hat{\beta}_0) + x^2 \hat{Var}(\hat{\beta}_1) + 2x \hat{Cov}(\hat{\beta}_0, \hat{\beta}_1)$$

- A 100(1-a)% Confidence Interval is thus obtained as

$$\hat{g}(x) \pm z_{1-\alpha/2} \hat{SE}(\hat{g}(x))$$

```
#Obtaining the Confidence Intervals for the Logit Function
logit_50 = schdglm$coefficients[,1][1] + 50*schdglm$coefficients[,1][2]
var_logit_50 = schdglm$coefficients[,2][1]^2 + (50^2)*schdglm$coefficients[,2][2]^2 +
2*50*schdglm$cov.unscaled[1,2]
logit_50; var_logit_50
  0.2366037
  0.06466011
```

The Multiple Logistic Regression Model

- The link model here is

$$g(x)=\beta_0+\beta_1 x_1+\ldots+\beta_p x_p$$
$$=\beta' x$$

And the *multiple logistic regression model* is consequently

$$\pi(x)=\frac{e^{(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)}}{1+e^{(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)}}$$
$$=\frac{e^{\beta' x}}{1+e^{\beta' x}}$$

- The likelihood function can be written as

$$\sum_{i=1}^{n}[y_i-\pi(x_i)]=0$$
$$\sum_{i=1}^{n}x_{ij}[y_i-\pi(x_i)]=0$$

Example: The Low Birth Weight Study

- Age
- LWT
- Race
- FTV: Number of First Trimester Physician Visit
- Low Birth Weight (LOW), the regressand

- The variable Race is divided into three variables (nominal)

```
lowbwt = read.xls("lowbwt.xls",sheet=1,header=T)
race2=race3=c()
for(i in 1:nrow(lowbwt))  {
if(lowbwt$RACE[i]==1)
{race2[i]=0;race3[i]=0}
if(lowbwt$RACE[i]==2)
{race2[i]=1;race3[i]=0}
if(lowbwt$RACE[i]==3)
{race2[i]=0;race3[i]=1}
}

lowglm=glm(LOW~AGE+LWT+race2+race3+FTV,data=lowbwt,family='binomial')
# Calculating the Log-Likelihood Value for the Full Model
design=cbind(rep(1,nrow(lowbwt)),lowbwt[,3],lowbwt[,4],race2,race3,lowbwt[,10])
logitx=design%*%lowglm$coefficients
pi_x=exp(logitx)/(1+exp(logitx))
fullloglik=sum((lowbwt$LOW)*log(pi_x))+sum((1-lowbwt$LOW)*log(1-pi_x))
```

```
summary(lowglm)
Call:
glm(formula = LOW ~ AGE + LWT + race2 + race3 + FTV, family = "binomial",
    data = lowbwt)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4163  -0.8931  -0.7113    1.2454   2.0755
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.295366   1.071443   1.209   0.2267
AGE         -0.023823   0.033730  -0.706   0.4800
LWT         -0.014245   0.006541  -2.178   0.0294 *
race2        1.003898   0.497859   2.016   0.0438 *
race3        0.433108   0.362240   1.196   0.2318
FTV         -0.049308   0.167239  -0.295   0.7681
---
Signif. codes:  0 â***â 0.001 â**â 0.01 â*â 0.05 â.â 0.1 â â 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 222.57  on 183  degrees of freedom
AIC: 234.57
```

Model Assessment

- We require two matrices once the regression coefficient estimates are available
- The *design matrix* **X**

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

- Define the matrix

$$V = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

- The observed *information matrix* can then be computed by

$$\hat{I}(\hat{\beta}) = X'VX$$

Calculating the log-likelihood values

-

```
design=cbind(rep(1,nrow(lowbwt)),lowbwt[,3],lowbwt[,4],race2,race3,lowbwt[,10])
low=lowbwt[,2]
logitx=design%*%lowglm$coefficients
pi_x=exp(logitx)/(1+exp(logitx))
sum((lowbwt$LOW)*log(pi_x))+sum((1-lowbwt$LOW)*log(1-pi_x))
```

-111.2865

The *G*- statistic for the significance of the model is given by

$$G = -2\ln\left[\frac{(likelihood\ without\ the\ variable)}{(likelihood\ with\ the\ variable)}\right]$$

where the likelihood under the null hypothesis is

$$\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}$$

Calculating the Log-Likelihood Value for the Null Model

```
n=nrow(design)
n1=sum(lowbwt$LOW); n0=n-n1
nullloglik=n1*log(n1/n) + n0*log(n0/n)
nullloglik
[1] -117.336
G = -2*(nullloglik-fullloglik)
12.09909
1-pchisq(G,5)
[1] 0.03345496
```

The p-value is significant at 0.05 level leads us to reject the null hypothesis and conclude that at least one or even all of p variable effect is significantly different from zero. From the summary we table, we select only the significant variables and fit the model again and test for model assessment using the G statistics.

- We now remove the insignificant variables from the “lowglm” model, and fit a new model
- Test for the likelihood ratio between the new model and the earlier model

```
lowglmfit=glm(LOW~LWT+race2+race3,data=lowbwt,family='binomial')
fitdesign=cbind(rep(1,nrow(lowbwt)),lowbwt[,4],race2,race3)
fitlogitx=fitdesign%*%lowglmfit$coefficients
fitpi_x=exp(fitlogitx)/(1+exp(fitlogitx))
fullloglikfit=sum((lowbwt$LOW)*log(fitpi_x))+sum((1-lowbwt$LOW)*log(1-fitpi_x))
Gfit = -2*(fullloglikfit-fullloglik)
1-pchisq(Gfit,2)
[1] 0.7095729
```

Thus, the two models are not significantly different, and we have not lost much by removing the insignificant variables from the full model

Confidence Intervals for Multiple Logistic Regression Model

```
glmconfint= function(glm,alpha)  {
glm=glm
n=length(glm$coefficients)
sdd=as.numeric(summary(glm)$coefficients[,2])
xmat=matrix(nrow=n, ncol=2)
colnames(xmat)=c("LCL","UCL")
rownames(xmat)= names(lowglm$coefficients)
xmat[,1]=as.numeric(glm$coefficients) - qnorm(1-alpha/2)*sdd
xmat[,2]=as.numeric(glm$coefficients) + qnorm(1-alpha/2)*sdd
return(xmat)
}
glmconfint(lowglm,0.05)
                  LCL         UCL
(Intercept) -0.80462420  3.395355695
AGE         -0.08993181  0.042285866
LWT         -0.02706418 -0.001425038
race2        0.02811220  1.979683451
race3       -0.27686939  1.143086244
FTV         -0.37709090  0.278474261
```

Estimation of Logit Function in Multiple Logistic Regression Model

- The link model may be estimated by

$$\hat{g}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p$$

$$= \hat{\boldsymbol{\beta}}' \boldsymbol{x}$$

and its estimated variance is

$$\hat{Var}[\hat{g}(\boldsymbol{x})] = \sum_{j=0}^{p} x_j^2 \hat{Var}(\hat{\beta}_j) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2 x_j x_k \hat{Cov}(\hat{\beta}_j, \hat{\beta}_k)$$

$$= \boldsymbol{x}' (X' V X)^{-1} \boldsymbol{x}$$

- There is a simple "trick" that can be used with the list object *glm*
- Ensures that we don't have to write long codes/program

Example. Estimating the logit of a 150 pound white women (for the reduced logistic regression model)

```
# R Function for estimating the logit and the variance
logitglm = function(glm,x)      {
x=c(1,x)
varmat=summary(glm)$cov.scaled
logitx= glm$coefficients%*%x
varlogitx=t(x)%*%varmat%*%x
xx=c(logitx,varlogitx)
names(xx)=c("Estimated Logit", "Variance of Estimated Logit")
return(xx)
}
#Estimating the Logit Function for a 150 pound white women
ww=c(150,0,0)
logitglm(lowglmfit,ww)
```

Estimated Logit	Variance of Estimated Logit
-1.4777	0.0832

Model-Building Strategies

- For an excellent discussion of variable selection, see Section 2, Chapter 4 of Hosmer and Lemeshow (2000).
- We discuss two methods of variable selection:
  - Stepwise Logistic Regression
  - Best Subsets Logistic Regression
- The approach is similar as in linear models

Stepwise Logistic Regression

- The most important variable is the one that produces maximum change in the log-likelihood relative to a model containing no variables
- That is, maximum change in the value of the *G*-statistic

Step 0.

- Consider all plausible variables, say they number to p
- Fit "intercept only model", and evaluate the log-likelihood L_0.
- Fit each of the possible *p* univariate logistic regression models and note the log-likelihood values $L_j^{(0)}$. Further, calculate the *G*-statistic for each of the *p*-models, $G_j^{(0)} = -2(L_0 - L_j^{(0)})$
- Obtain the *p-value* for each model
- The most important variable is then the one with least *p*-value, $p_{e1}^{(0)} = min(p_j^{(0)})$

Stepwise Logistic Regression

- Denote the most important variable by x_{e1}
- Define the entry criteria p-value as p_E, which will at any time throughout this procedure decide if the variable is to be included or not. That is, the variable with least p-value must be lesser than p_E to be selected in the final model
- If none of the variables have the p-value lesser than p_E, we stop

Step 1. The Forward Selection Step

- Replace L_0 of the previous step with $L_{e1}^{(1)}$
- Fit p-1 models with variables x_{e1} and the remaining variables x_j, $j = 1, 2, ..., p$, and j distinct from e_1.
- For each of the p-1 models, calculate the log-likelihood $L_{e1,j}^{(1)}$ and the G-statistics $G_j^{(1)} = -2(L_{e1}^{(1)} - L_{e1,j}^{(1)})$, and the corresponding p-values, denoted $p_j^{(1)}$

Stepwise Logistic Regression

- Define $p_{e2}^{(1)} = min(p_j^{(1)})$
- If $p_{e2}^{(1)} = p_E$, stop

Step 2. The Backward Elimination Step

- Adding x_{e2} may leave x_{e1} statistically insignificant.
- Let $L_{-ej}^{(2)}$ denote the log-likelihood of the model with variable e_j removed
- Calculate the likelihood-ratio test of these reduced models with respect to the full model at the beginning of this step $G_{-ej}^{(2)} = -2(L_{ej}^{(2)} - L_{e1e2}^{(2)})$, and calculate the p-values $p_{-ej}^{(2)}$
- Deleted variables must result in a maximum p-value of the modified model
- Denote x_{r2} as the variable which is to be removed, and define $p_{r2}^{(2)} = max(p_{-e1}^{(2)}, p_{-e2}^{(2)})$

Stepwise Logistic Regression

- To remove variables, we need to have a value p_R with respect to which we compare $p_{r2}^{(2)}$
- We need to have $p_R > p_E$ **(WHY?)**
- Variables are removed if $p_{r2}^{(2)} > p_R$

Forward Selection Phase.

- Continue the forward selection method with *p-2* remaining variables and find $p_{e3}^{(2)}$
- Let x_{e3} be the variable associated with $p_{e3}^{(2)}$.
- If the $p_{e3}^{(2)}$ is lesser than p_E, proceed to Step 3, otherwise stop.

Stepwise Logistic Regression

Step 3.

- Fit the model including the variable selected in the previous step and perform backward elimination and then forward selection phase
- Repeat untill the last Step S

Step (S)

- Stopping happens if all the variables have been selected in the model
- Also happens if all the p-values in the model are less than p_R, and the remaining variables have p-values exceed p_E.

Stepwise Logistic Regression

- The "step" function in R may be used to build model
- The criteria used there is "AIC" though
- To the best of our knowledge, there is no package/function which will implement stepwise logistic regression using the G-statistic generated p-values
- As we are following Hosmer and Lemeshow (2000) from the beginning of the presentation, we will write exhaustive codes to illustrate the the stepwise logistic regression
- We continue the example of Low Birth Weight study
- Towards this, we first define two important functions
  - glmllv : Given a "glm" model, the output y, and the covariate vector, this function returns the loglikelihood value
  - pvalue: Given a loglikelihood values of base model and current model and the degrees of freedom, we get the p-value

glmllv and pvalue Functions

```
glmllv = function(glm, x)      {
if(is.vector(x)==T) x1=cbind(rep(1,length(x)),x)
if(is.matrix(x)==T) x1=cbind(rep(1,nrow(x)),x)
coeff=as.numeric(glm$coefficients)
y=as.numeric(glm$y)
some=x1%*%coeff; logit=exp(some)/(1+exp(some))
llvalue=sum(y*logit)+sum((1-y)*log(logit))
return(llvalue)
}
pvalue = function(lik1,lik0,df)     {
gstat=-2*(lik0-lik1)
pval=1-pchisq(gstat,df)
return(pval)
}
```

Stepwise Logistic Regression for the Low Birth Weight Study

- We first calculate the null-likelihood value and the *p*- univariate logistic regression values and note that minimum *p*-value

```
library(gdata)
lowbwt = read.xls("lowbwt.xls",sheet=1,header=T)
attach(lowbwt)
RACE_2=RACE_3=c()
for(i in 1:nrow(lowbwt))   {
if(lowbwt$RACE[i]==1) {RACE_2[i]=0;RACE_3[i]=0}
if(lowbwt$RACE[i]==2) {RACE_2[i]=1;RACE_3[i]=0}
if(lowbwt$RACE[i]==3) {RACE_2[i]=0;RACE_3[i]=1}
}
design=cbind(rep(1,nrow(lowbwt)),lowbwt[,3],lowbwt[,4],RACE_2,RACE_3,lowb
wt[,10])
colnames(design)=c("intercept", "AGE","LWT","RACE_2","RACE_3","FTV")
n=nrow(design)
n1=sum(lowbwt$LOW); n0=n-n1
nullloglik=n1*log(n1/n) + n0*log(n0/n)
```

[1] -117.336

Low Birth Weight Study: Step 0

```
#The p-values for entry and exit criteria
pe=0.25; pr=0.4
# Selecting the first variable to be included in the model
glm_AGE=glm(LOW~AGE,family='binomial')
ll_AGE=glmllv(glm_AGE,AGE)
(pvalue_AGE=pvalue(ll_AGE,nullloglik,1))
[1] 0.09664596
glm_LWT=glm(LOW~LWT,family='binomial')
ll_LWT=glmllv(glm_LWT,LWT)
(pvalue_LWT=pvalue(ll_LWT,nullloglik,1))
[1] 0.01445812
glm_RACE_2=glm(LOW~RACE_2,family='binomial')
ll_RACE_2=glmllv(glm_RACE_2,RACE_2)
(pvalue_RACE_2=pvalue(ll_RACE_2,nullloglik,1))
[1] 0.1985105
glm_RACE_3=glm(LOW~RACE_3,family='binomial')
ll_RACE_3=glmllv(glm_RACE_3,RACE_3)
(pvalue_RACE_3=pvalue(ll_RACE_3,nullloglik,1))
[1] 0.1829021
glm_FTV=glm(LOW~FTV,family='binomial')
ll_FTV=glmllv(glm_FTV,FTV)
(pvalue_FTV=pvalue(ll_FTV,nullloglik,1))
[1] 0.3792461
```

We see that the minimum p-value of 0.0145 is associated with LWT variable, and its also lesser than p_E. We include this variable in the model now.

Low Birth Weight Study: Step 1

```
#Selecting the variables for the Step 1
glm_base=glm(LOW~LWT,family='binomial')
ll_base=glmllv(glm_base,LWT)
glm_base_AGE=glm(LOW~LWT+AGE,family='binomial')
ll_base_AGE=glmllv(glm_base_AGE,cbind(LWT,AGE))
(pvalue_base_AGE=pvalue(ll_base_AGE,ll_base,1))
[1] 0.2106024
glm_base_RACE_2=glm(LOW~LWT+RACE_2,family='binomial')
ll_base_RACE_2=glmllv(glm_base_RACE_2,cbind(LWT,RACE_2))
(pvalue_base_RACE_2=pvalue(ll_base_RACE_2,ll_base,1))
[1] 0.05723459
glm_base_RACE_3=glm(LOW~LWT+RACE_3,family='binomial')
ll_base_RACE_3=glmllv(glm_base_RACE_3,cbind(LWT,RACE_3))
(pvalue_base_RACE_3=pvalue(ll_base_RACE_3,ll_base,1))
[1] 0.442516
glm_base_FTV=glm(LOW~LWT+FTV,family='binomial')
ll_base_FTV=glmllv(glm_base_FTV,cbind(LWT,FTV))
(pvalue_base_FTV=pvalue(ll_base_FTV,ll_base,1))
[1] 0.5457832
```

Since the p-value associated with RACE_2 is least and it is lesser than p_E, it can be now selected in our model

Low Birth Weight Study: Step 2

- We need to now test if we remove LWT or RACE from the model, does the model become less significant.

```
#Backward Elimination Method of Step 2
# Since Race is consists of both RACE_2 and RACE_3, we include both in Step 2
glm_base_RACE=glm(LOW~LWT+RACE_2+RACE_3,family='binomial')
ll_base_RACE=glmllv(glm_base_RACE, cbind(LWT,RACE_2,RACE_3))
glm_RACE=glm(LOW~RACE_2+RACE_3)
ll_RACE=glmllv(glm_RACE,cbind(RACE_2,RACE_3))
pvalue(ll_base_RACE,ll_LWT,2)
[1] 0.06615272
pvalue(ll_base_RACE,ll_RACE,1)
[1] 1.554312e-15
```

Since the maximum of these two p-values is lesser than p_R, we retain the variable RACE in the model. That is, the backward elimination step has not removed any variable.

Low Birth Weight Study: Step 3

- We need to redo the Step 2 untill we reach the Step S described earlier.

```
#Step 3 continues the Step 2 untill stopping criteria
glm_LWT_RACE_AGE=glm(LOW~LWT+RACE_2+RACE_3+AGE)
ll_LWT_RACE_AGE=glmllv(glm_LWT_RACE_AGE,cbind(LWT,RACE_2,RACE_3,AGE))
(pvalue_LWT_RACE_AGE=pvalue(ll_LWT_RACE_AGE,ll_LWT_RACE,1))
[1] 1
glm_LWT_RACE_FTV=glm(LOW~LWT+RACE_2+RACE_3+FTV)
ll_LWT_RACE_FTV=glmllv(glm_LWT_RACE_FTV,cbind(LWT,RACE_2,RACE_3,FTV))
(pvalue_LWT_RACE_FTV=pvalue(ll_LWT_RACE_FTV,ll_LWT_RACE,1))
[1] 1
```

Since none of the p-values associated variables AGE and FTV is less than p_E, we can't enter the variables into the model. Thus, our best model includes the variables LTW and RACE.

The Probit Regression Model

- Recall that in the logistic regression we had

$$P(Y=1|x)=\pi(x)=\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$

$$P(Y=0|x)=\pi(x)=\frac{1}{1+e^{(\beta_0+\beta_1 x)}}$$

- We essentially ensured that the quantity on the right hand side is always positive
- *Probit regression* model uses the normal cumulative distribution on the RHS to ensure the same thing. Thus, we have the model

$$P(Y=1|x)=\Phi(\beta_0+\beta_1 x)$$

$$P(Y=0|x)=1-\Phi(\beta_0+\beta_1 x)$$

- We will not go in much details and close the discussion with the simple exmaple of Coronarry Heart Disease

```
chdage=read.xls("chdage.xls",sheet=1,header=T)
chdprobit=glm(chdage$CHD~chdage$AGE,binomial(probit))
summary(chdprobit)
Call:
glm(formula = chdage$CHD ~ chdage$AGE, family = binomial(probit))
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9713  -0.8608  -0.4499   0.8359   2.3269
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.14573    0.62460  -5.036 4.74e-07 ***
chdage$AGE   0.06580    0.01335   4.930 8.20e-07 ***
---
Signif. codes:  0 â***â 0.001 â**â 0.01 â*â 0.05 â.â 0.1 â â 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 136.66  on 99  degrees of freedom
Residual deviance: 107.50  on 98  degrees of freedom
AIC: 111.50
Number of Fisher Scoring iterations: 4
```

POISSON REGRESSION

- The Poisson distribution is

$$P(Y=y)=\frac{e^{-\lambda}\lambda^{y}}{y!},\ y=0,1,2,\ldots$$

- We further know that

$$E(Y)=\lambda \text{ and } Var(Y)=\lambda$$