Statistical Modeling
BANA 7042

Lecture 2: Logistic regression model

What might cause heart disease?

Western Collaborative Group Study

- One of the earliest studies that addressed the heart disease issue.
- Started in 1960.
- 3154 healthy men:
  - Aged from 39 to 59;
  - From the San Francisco area;
  - Free of heart disease at the start of the study.
- 8.5 years later, the study recorded whether these men suffered from heart disease along with many other variables that might be related.
- Rosenman et al. (1975), JAMA, 233(8), 872-877.

Load the data set “wcgs”

```
install.packages("faraway")
library("faraway")

data(wcgs)
class(wcgs)
str(wcgs)
```

The data structure

```
> str(wcgs)
'data.frame':   3154 obs. of  13 variables:
 $ age    : int  49 42 42 41 59 44 44 40 43 42 ...
 $ height : int  73 70 69 68 70 72 72 71 72 70 ...
 $ weight : int  150 160 160 152 150 204 164 150 190 175 ...
 $ sdp    : int  110 154 110 124 144 150 130 138 146 132 ...
 $ dbp    : int  76 84 78 78 86 90 84 60 76 90 ...
 $ chol   : int  225 177 181 132 255 182 155 140 149 325 ...
 $ behave : Factor w/ 4 levels "A1","A2","B3",..: 2 2 3 4 3 4 4 2 3 2 ...
 $ cigs   : int  25 20 0 20 20 0 0 0 25 0 ...
 $ dibep  : Factor w/ 2 levels "A","B": 2 2 1 1 1 1 1 2 1 2 ...
 $ chd    : Factor w/ 2 levels "no","yes": 1 1 1 1 2 1 1 1 1 1 ...
 $ typechd: Factor w/ 4 levels "angina","infdeath",..: 3 3 3 3 2 3 3 3 3 3 ...
 $ timechd: int  1664 3071 3071 3064 1885 3102 3074 3071 3064 1032 ...
 $ arcus  : Factor w/ 2 levels "absent","present": 1 2 1 1 2 1 1 1 1 2 ...
```

The meaning of each variable?

```
?wcgs
```

age

age in years

height

height in inches

weight

weight in pounds

sdp

systolic blood pressure in mm Hg

dbp

diastolic blood pressure in mm Hg

chol

Fasting serum cholesterol in mm %

behave

behavior type which is a factor with levels A1 A2 B3 B4

cigs

number of cigarettes smoked per day

dibep

behavior type a factor with levels A (Agressive) B (Passive)

chd

coronary heat disease developed is a factor with levels no yes

typechd

type of coronary heart disease is a factor with levels angina infdeath none silent

timechd

Time of CHD event or end of follow-up

arcus

arcus senilis is a factor with levels absent present

Focusing on 3 variables

```
summary(wcgs[,c("chd","height","cigs")])
 chd             height           cigs      
no :2897   Min.   :60.00   Min.   : 0.0  
yes: 257   1st Qu.:68.00   1st Qu.: 0.0  
           Median :70.00   Median : 0.0  
           Mean   :69.78   Mean   :11.6  
           3rd Qu.:72.00   3rd Qu.:20.0  
           Max.   :78.00   Max.   :99.0  
```

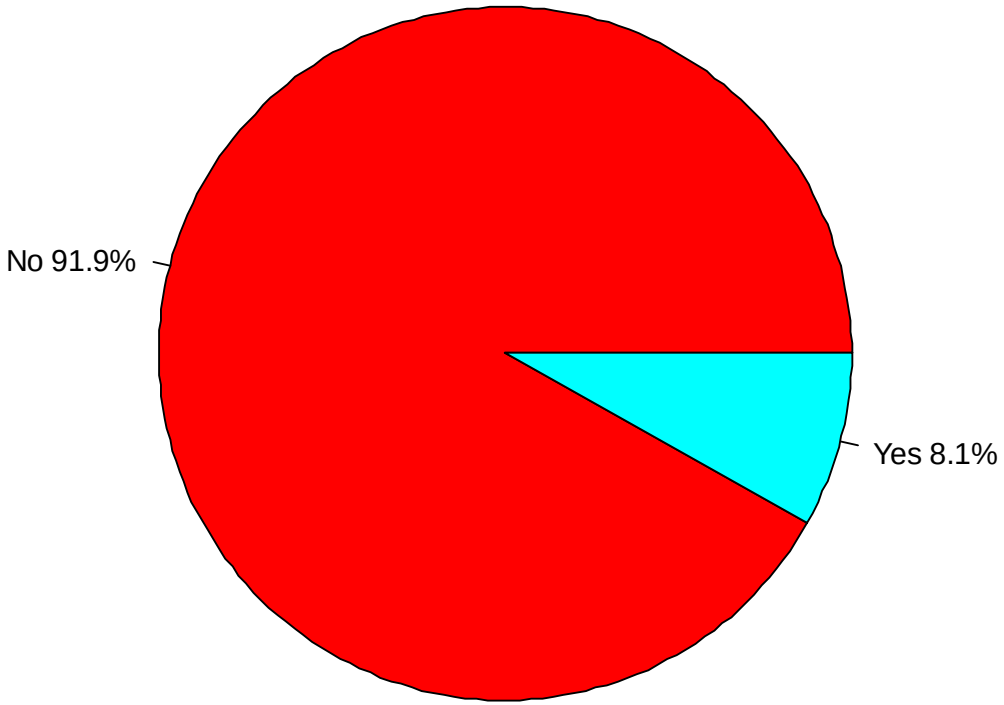
Pay attention to these statistics!

```
summary(wcgs[,c("chd","height","cigs")])
  chd             height           cigs      
 no :2897   Min.   :60.00   Min.   : 0.0  
 yes: 257   1st Qu.:68.00   1st Qu.: 0.0  
            Median :70.00   Median : 0.0  
            Mean   :69.78   Mean   :11.6  
            3rd Qu.:72.00   3rd Qu.:20.0  
            Max.   :78.00   Max.   :99.0  
```

Pie chart of the binary response

```
attach(wcgs)
pct<-round(table(chd)/length(chd)*100,1)
labs<-c("No","Yes")
labs<-paste(labs,pct)
labs<-paste(labs,"%",sep="" )
pie(table(chd),labels=labs,col=rainbow(length(labs)),m
ain="Pie chart of Coronary Heart Disease")
```

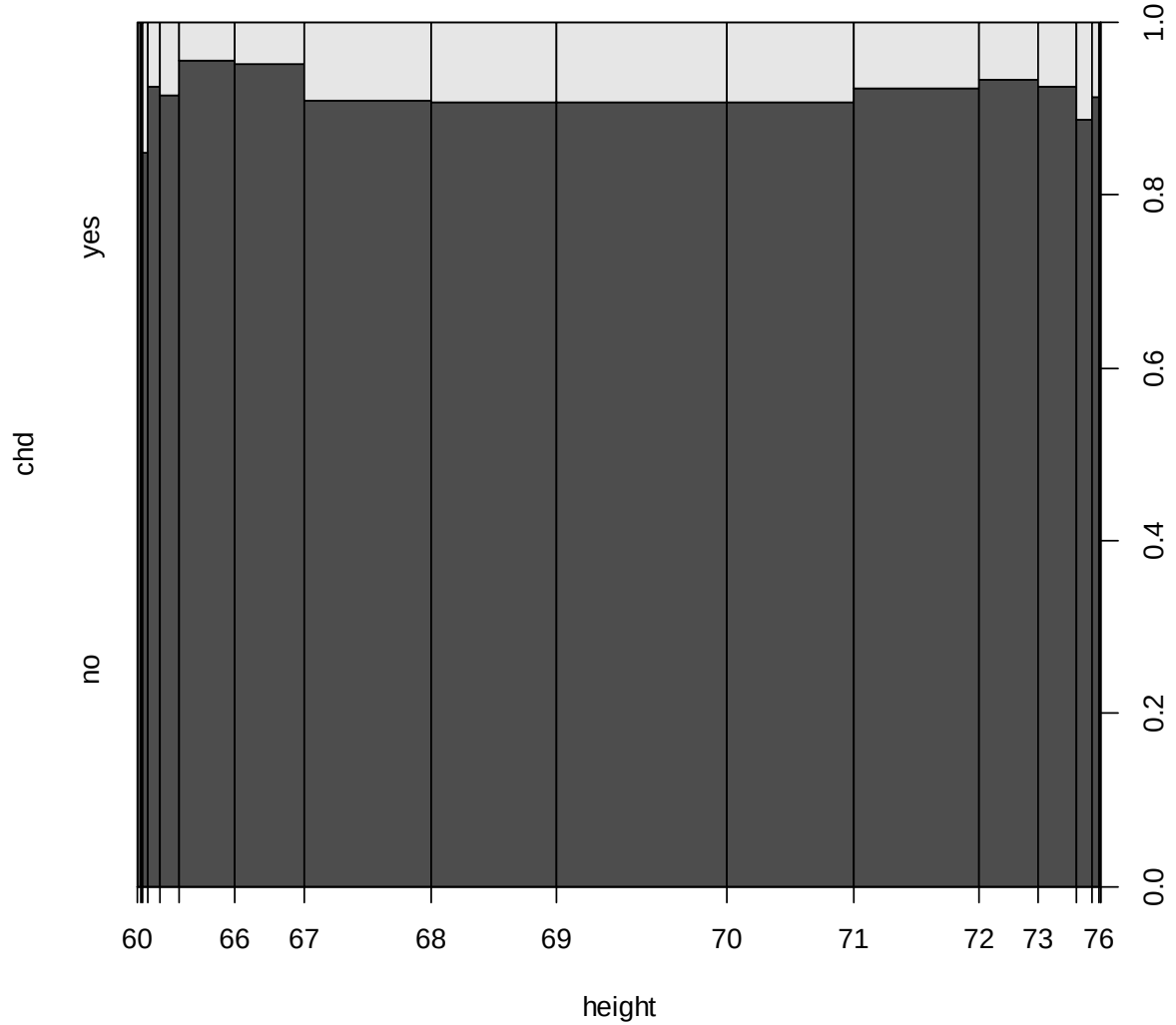
Pie chart of Coronary Heart Disease

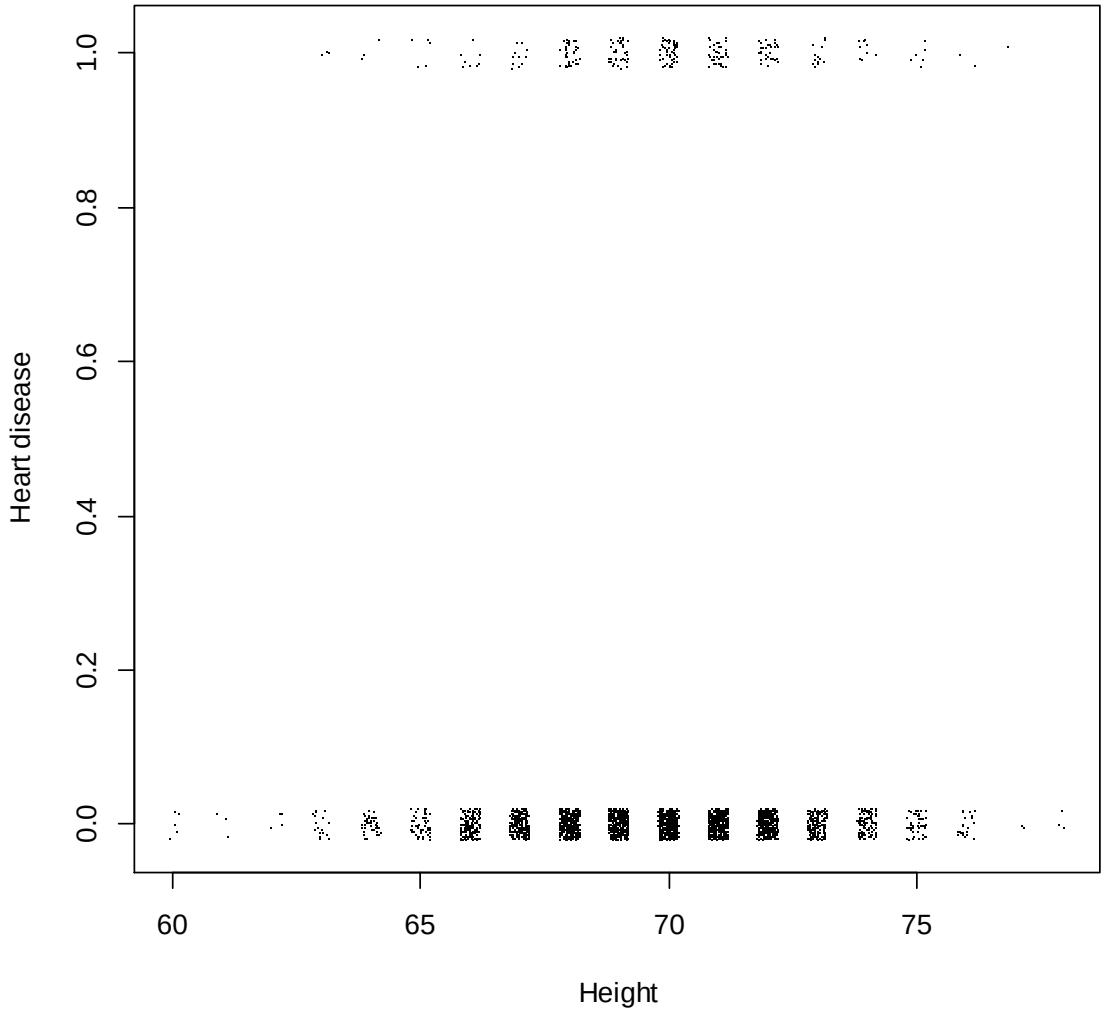


Visualize the association

- **"chd" versus "height"**

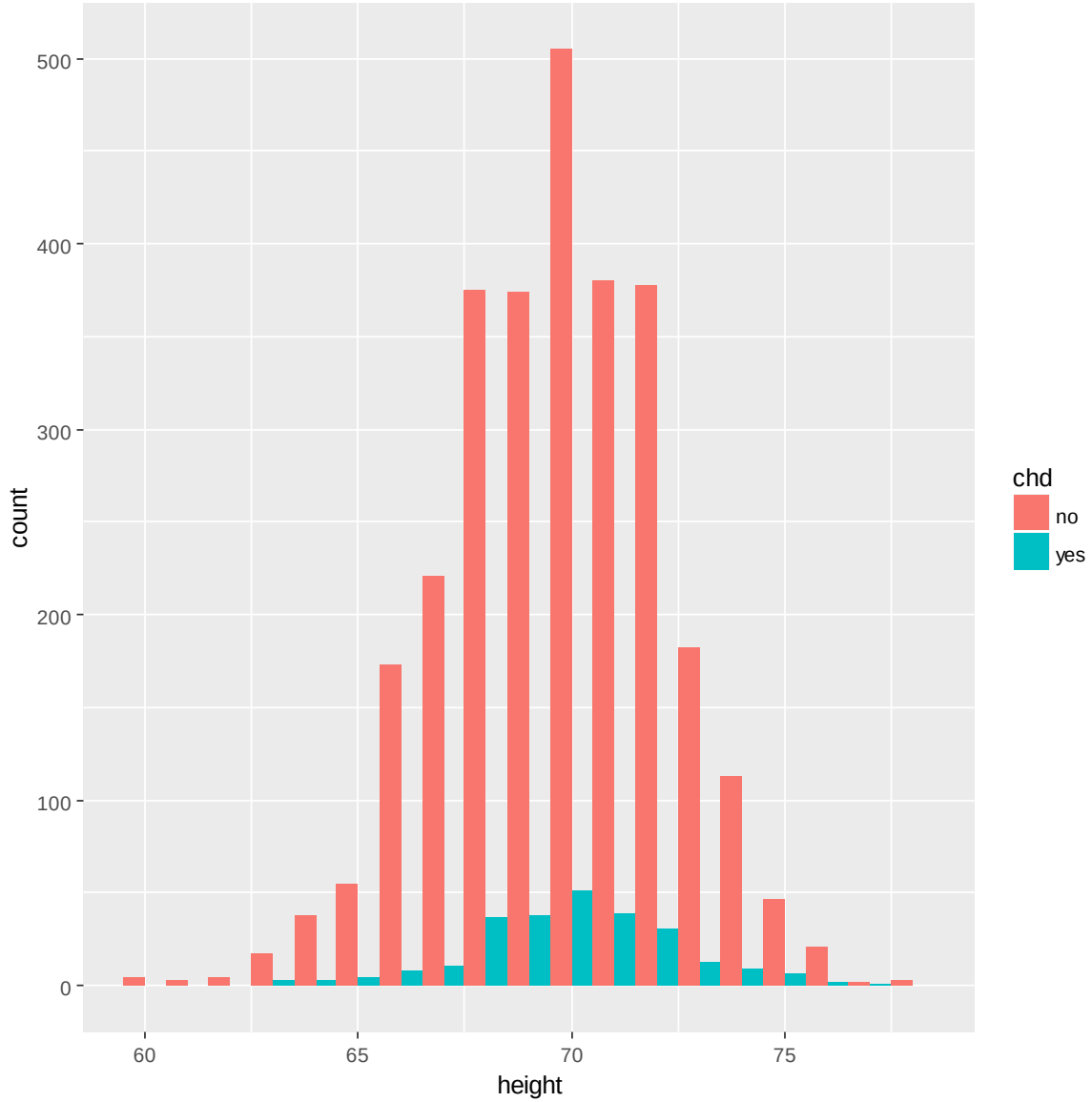
```
plot(chd~height)


wcgs$y<-ifelse(chd=="no",0,1)
windows()
plot(jitter(y,0.1)~jitter(height),wcgs,xlab="Height",y
lab="Heart disease",pch=".")
```


chd
yes
no
1.0
0.8
0.6
0.4
0.2
0.0
60
66
67
68
69
70
71
72
73
76
height


Heart disease
0.0
0.2
0.4
0.6
0.8
1.0
60
65
70
75
Height

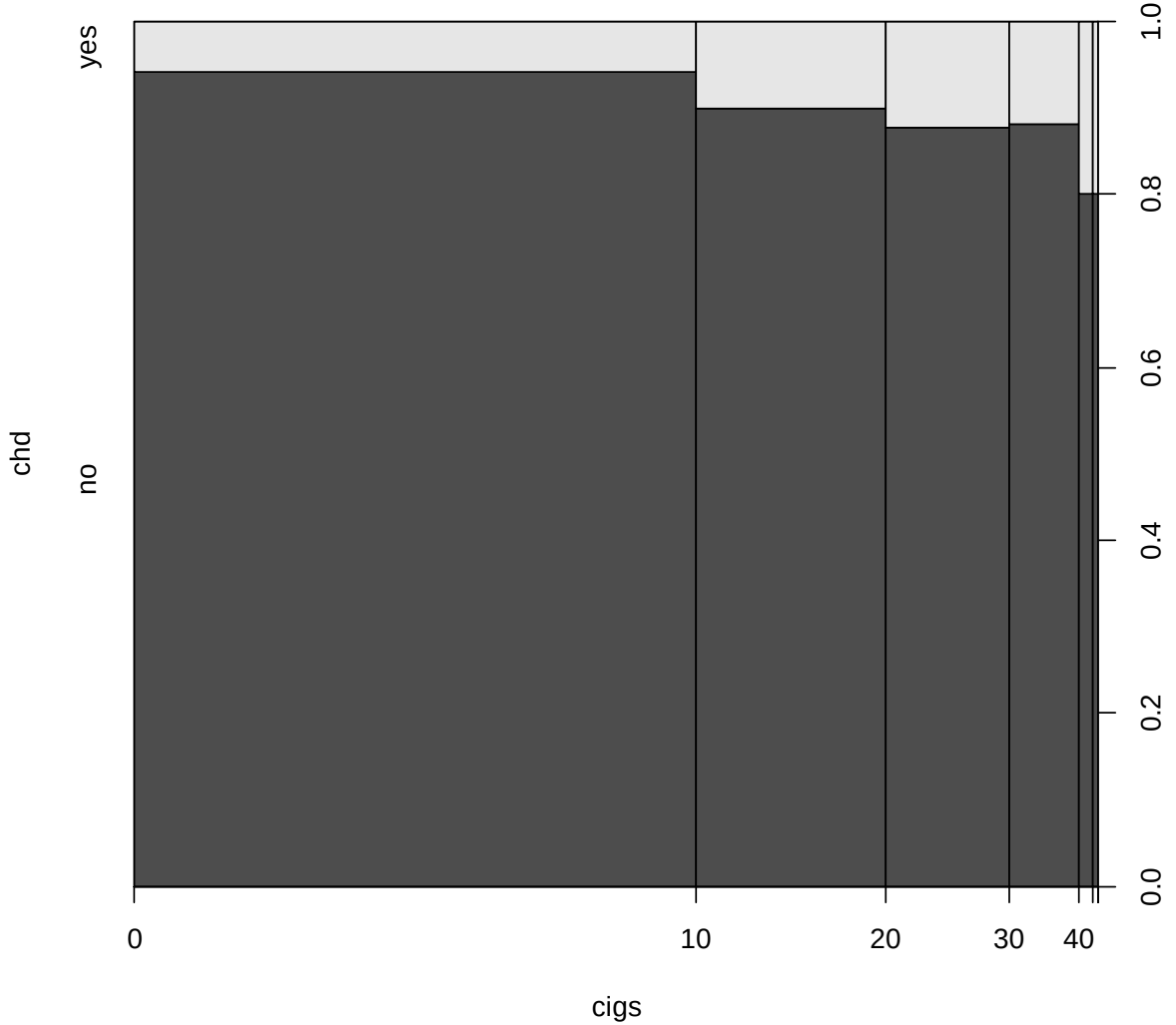
ggplot

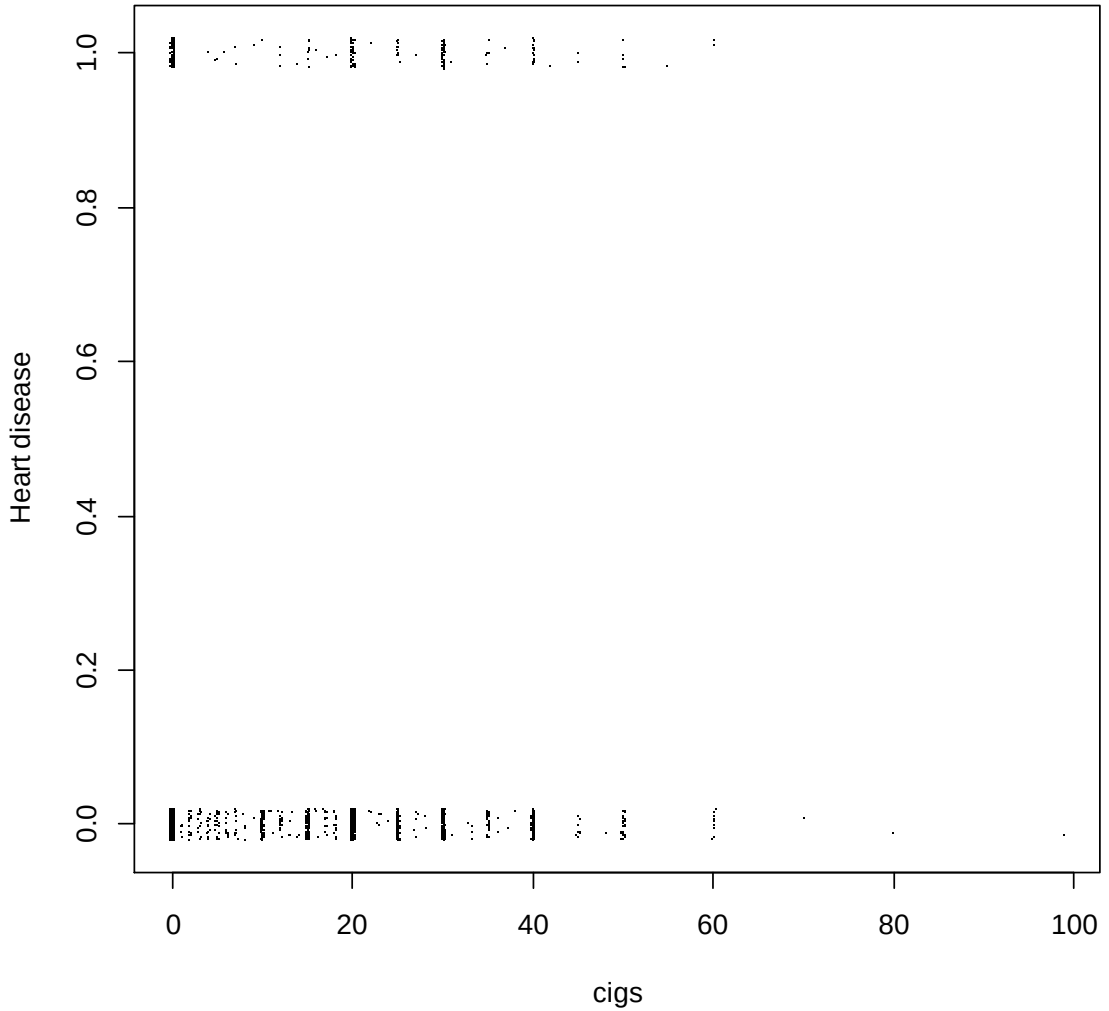
```
install.packages("ggplot2")
library("ggplot2")

ggplot(wcgs,aes(x=height,fill=chd))+geom_histogram(pos
ition="dodge",binwidth=1)
```

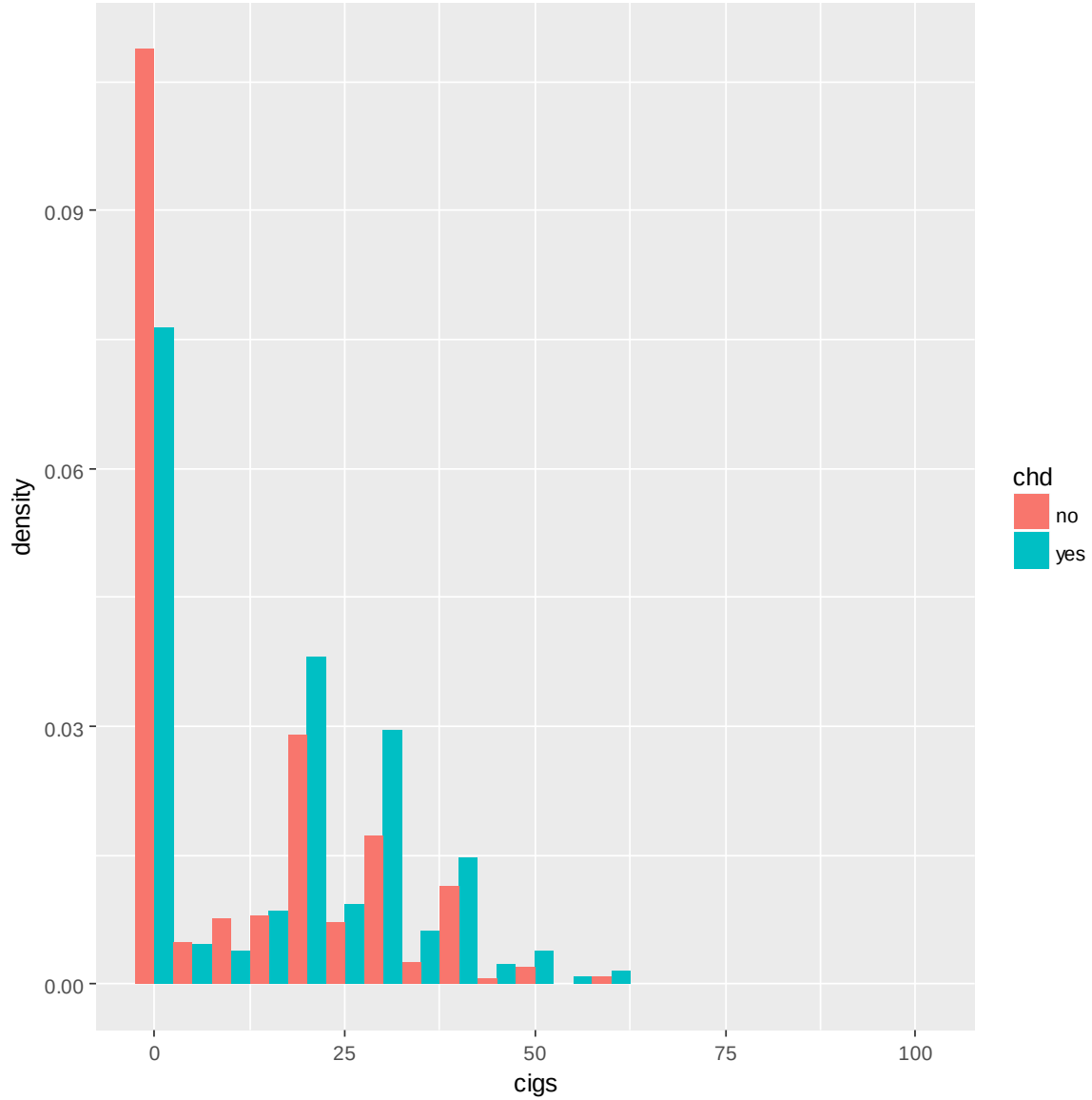

500
400
300
200
100
0
count
60
65
70
75
height
chd
no
yes

Can you visualize the association between “chd” and “cigs”?

```
### Visualize chd versus cigs
plot(chd~cigs)


windows()
plot(jitter(y,0.1)~jitter(cigs),wcgs,xlab="cigs",ylab=
"Heart disease",pch=".")


windows()
ggplot(wcgs,aes(x=cigs,fill=chd))+geom_histogram(posit
ion="dodge",binwidth=5,aes(y=..density..))
```


chd
yes
no
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
cigs


Heart disease
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
cigs


density
0.09
0.06
0.03
0.00
0
25
50
75
100
cigs
chd
no
yes

Our observations so far

- It seems that the variable “cigs” is positively associated with the binary response “chd”.
- It is not clear if “height” is associated with “chd” or not.
- Question: how can we build a model to examine the association?

Does the linear regression model work?

Please try.

```
model.lm<-lm(y~height+cigs,wcgs)
summary(model.lm)
```

Question: can we use this model to make inference?

```
Call:
lm(formula = y ~ height + cigs, data = wcgs)

Residuals:
     Min       1Q   Median       3Q      Max
-0.25454 -0.09831 -0.06298 -0.05736  0.95387

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0718275  0.1338592  -0.537    0.592
height       0.0018723  0.0019171   0.977    0.329
cigs         0.0019539  0.0003339   5.851 5.38e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2722 on 3151 degrees of freedom
Multiple R-squared:  0.0111,    Adjusted R-squared:  0.01047
F-statistic: 17.69 on 2 and 3151 DF,  p-value: 2.303e-08
```

Any problem with prediction?

- Using the fitted linear regression model, what is your predicted value of "chd" for a man whose height is 80 inches and who consumes 1000 cigarettes daily (we make an extreme case)?

```
beta<-coef(model.lm)
beta[1]+beta[2]*80+beta[3]*1000
2.031886
```

How can we model a binary response using a number of covariates?

Recall the case of a continuous response

Suppose the normality assumption holds

$$LD \sim N(\mu, \sigma^2)$$

How can we associate an explanatory variable (or predictor, covariate..) X to the response variable $Y = LD$?

Establish association through the mean

- We will make two additional assumptions on top of

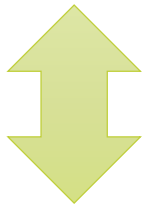
$$LD \sim N(\mu, \sigma^2)$$

1. LD relies on X only through its mean, e.g, $LD \sim N(\mu(X), \sigma^2)$.
2. Linearity: $\mu(X) = \beta_0 + X\beta_1$

- In other words, $LD \sim N(\beta_0 + X\beta_1, \sigma^2)$

It is a linear regression!

$$LD \sim N(\beta_0 + X\beta_1, \sigma^2)$$



$$LD = \beta_0 + X\beta_1 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

How can we extend the idea to the case of a binary response?

- Suppose the binary response

$$Y \sim \text{Bernoulli}(p)$$

- We will make two additional assumptions on top of this assumption:

How can we extend the idea to the case of a binary response?

- Suppose the binary response

$$Y \sim \text{Bernoulli}(p)$$

- We will make two additional assumptions on top of this assumption:
  1. Y relies on X only through its mean, e.g $Y \sim \text{Bernoulli}(p(X))$
  2. What is our second assumption?

$$p(X) = \beta_1 + \beta_2 X?$$

How can we extend the idea to the case of a binary response?

- Suppose the binary response

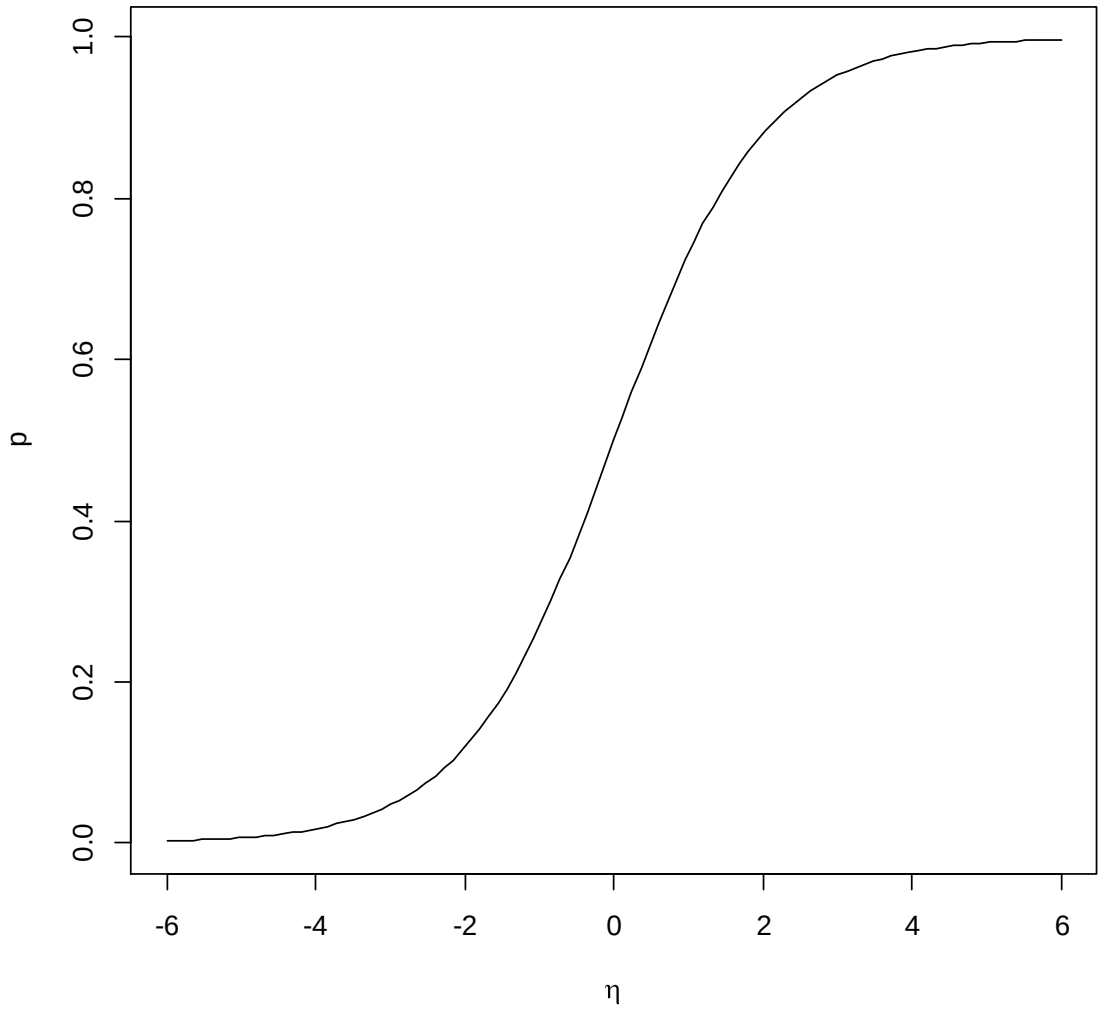
$$Y \sim \text{Bernoulli}(p)$$

- We will make two additional assumptions on top of this assumption:
  1. Y relies on X only through its mean, e.g $Y \sim \text{Bernoulli}(p(X))$
  2. The logit transformation of the parameter p

$$\text{logit}(p) = \log(\frac{p}{1-p}) = \beta_1 + \beta_2 X$$

$$\log(\frac{p}{1-p}) = \eta = \beta_1 + \beta_2 X$$

$$p = \frac{e^{\eta}}{1+e^{\eta}}$$



Fit a logistic regression model

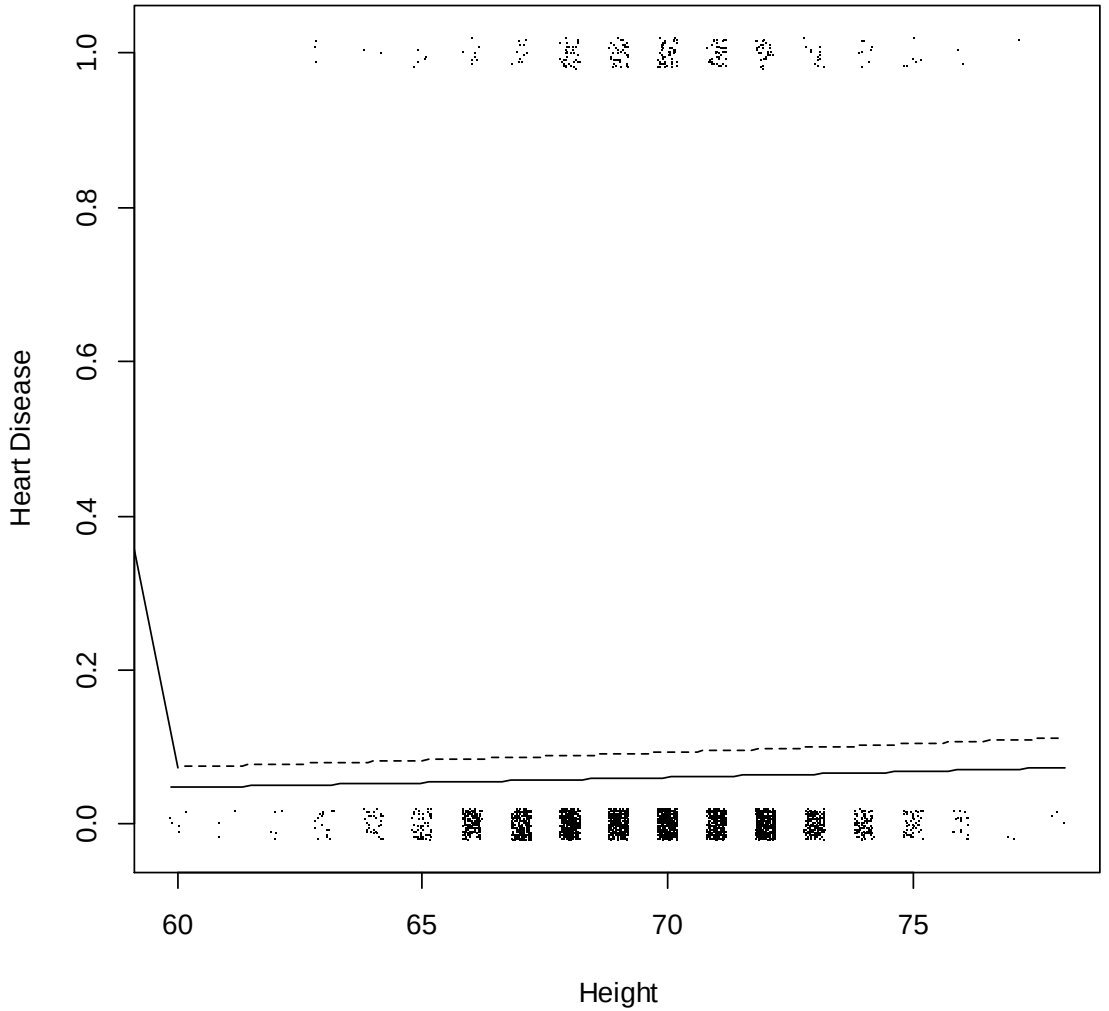
```
lmod<-glm(chd~height+cigs,family=binomial,wcgs)
summary(lmod)

Call:
glm(formula = chd ~ height + cigs, family = binomial, data = wcgs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0041  -0.4425  -0.3630  -0.3499   2.4357

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.50161    1.84186  -2.444   0.0145 *
height       0.02521    0.02633   0.957   0.3383
cigs         0.02313    0.00404   5.724 1.04e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1781.2  on 3153  degrees of freedom
Residual deviance: 1749.0  on 3151  degrees of freedom
AIC: 1755

Number of Fisher Scoring iterations: 5
```

Prediction

- We can compute the probability of heart disease given the values of the predictors.
- For example, varying the height for fixed levels of cigarette consumptions – nonsmokers and 20 cigarettes a day.

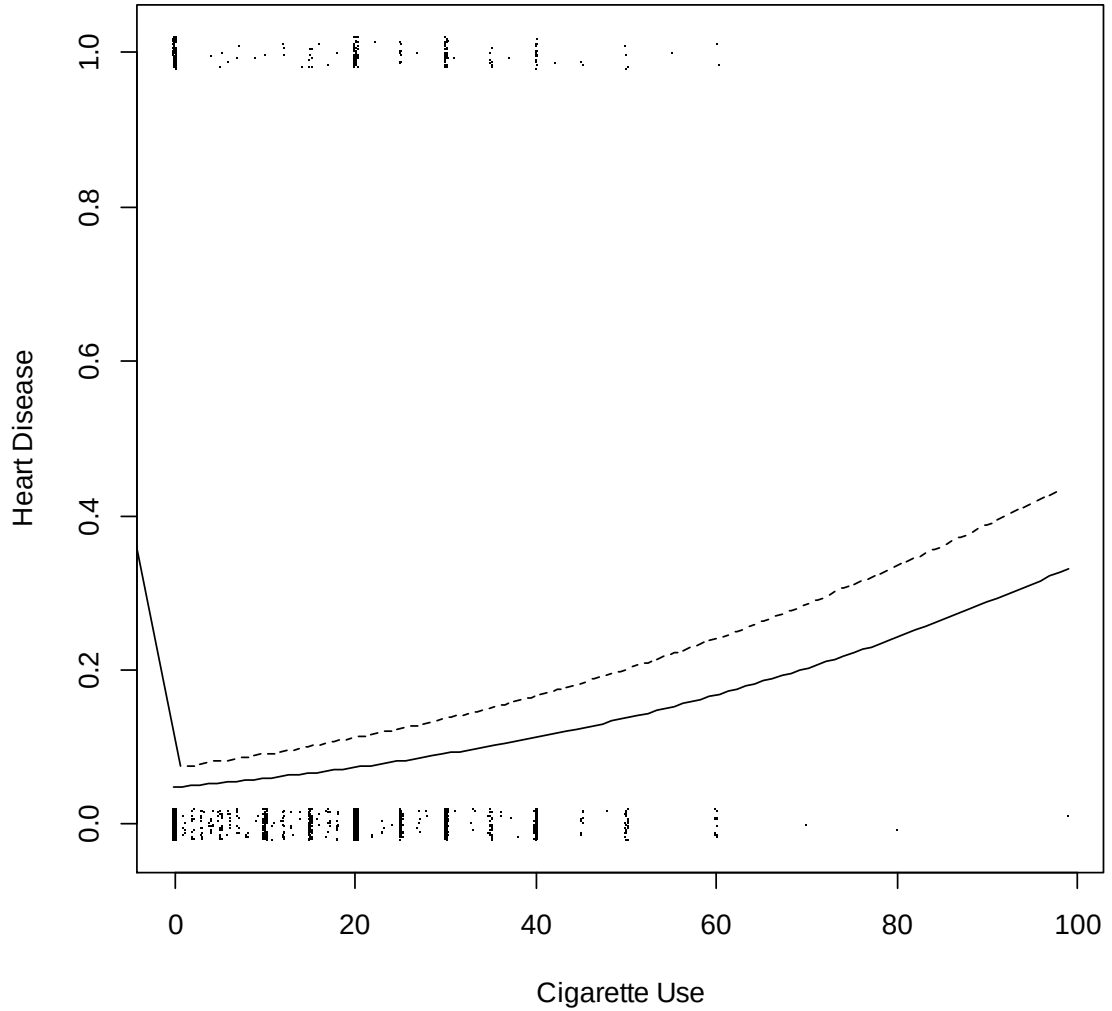
- R code

```
beta.lmod<-coef(lmod)
plot(jitter(y,0.1)~jitter(height),wcgs,xlab="Height",ylab="Heart
Disease",pch=".")
curve(ilogit(beta.lmod[1]+beta.lmod[2]*x+beta.lmod[3]*0),add=TRUE)
curve(ilogit(beta.lmod[1]+beta.lmod[2]*x+beta.lmod[3]*20),lty=2,add=TRUE)
```


Heart Disease
0.0
0.2
0.4
0.6
0.8
1.0
60
65
70
75
Height

```
plot(jitter(y,0.1)~jitter(cigs),wcgs,xlab="Cigarette Use",ylab="Heart Disease",pch=".")
curve(ilogit(beta.lmod[1]+beta.lmod[2]*60+beta.lmod[3]*x),add=TRUE)
curve(ilogit(beta.lmod[1]+beta.lmod[2]*78+beta.lmod[3]*x),lty=2,add=TRUE)
```

Question: how to interpret this plot?



Linear model vs logistic model

- Compare the linear model “model.lm” and the logistic model “lmod”. Do you think the linear model is useless?

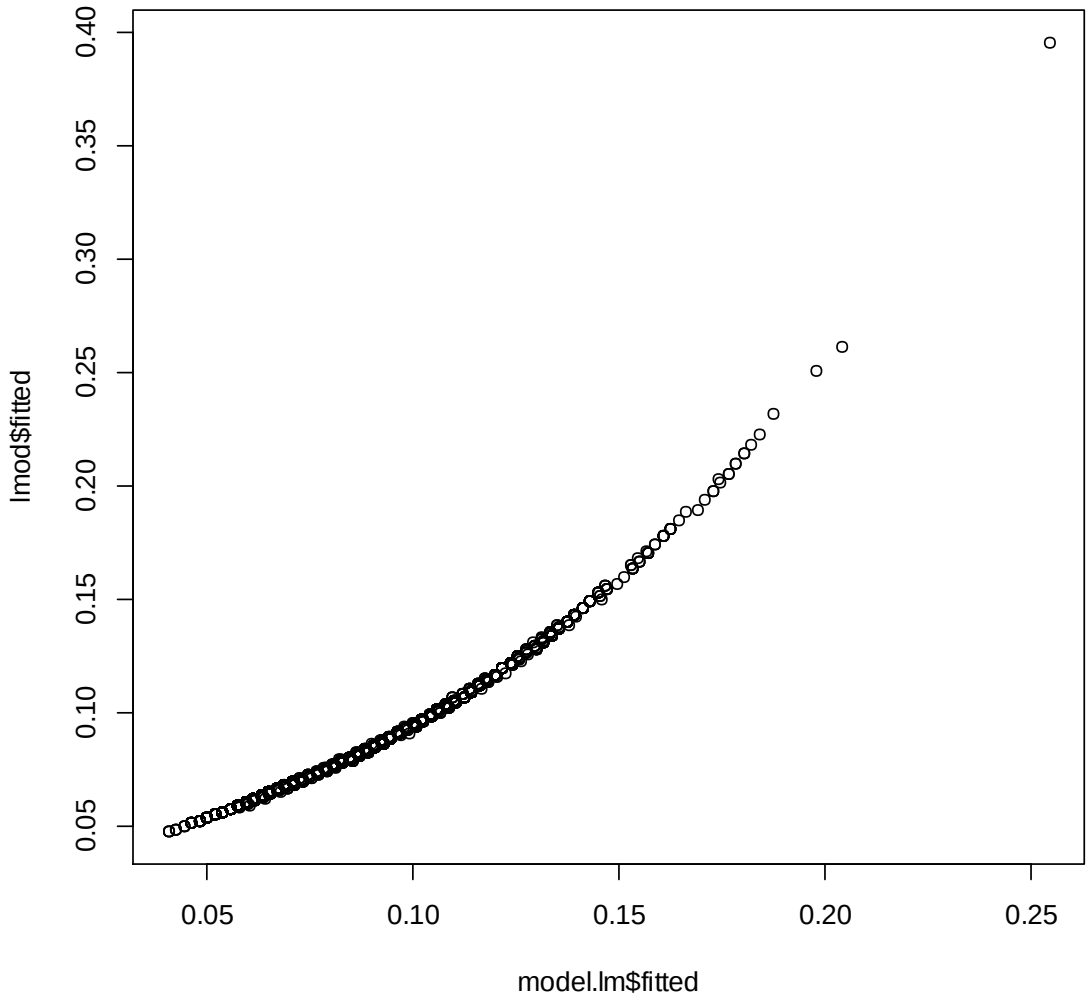
- Compare the following things:
  - The significance of the coefficients;
  - The signs of the coefficients;
  - The predicted values from the two models.

```
Linear model
lm(formula = y ~ height + cigs, data = wcgs)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0718275  0.1338592  -0.537    0.592
height       0.0018723  0.0019171   0.977    0.329
cigs         0.0019539  0.0003339   5.851 5.38e-09 ***

Logistic model
glm(formula = chd ~ height + cigs, family = binomial, data = wcgs)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.50161    1.84186  -2.444   0.0145 *
height       0.02521    0.02633   0.957   0.3383
cigs         0.02313    0.00404   5.724 1.04e-08 ***
```

Compare the predicted values

```
summary(model.lm$fitted)
summary(lmod$fitted)
plot(lmod$fitted~model.lm$fitted)
```

```
> summary(model.lm$fitted)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.04051 0.05923 0.06518 0.08148 0.10018 0.25454
> summary(lmod$fitted)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.04792 0.06082 0.06529 0.08148 0.09541 0.39598
```


lmod$fitted
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.05
0.10
0.15
0.20
0.25
model.lm$fitted

How to interpret regression coefficients?

```
glm(formula = chd ~ height + cigs, family = binomial, data
= wcgs)


Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.50161    1.84186  -2.444   0.0145 *
height       0.02521    0.02633   0.957   0.3383
cigs         0.02313    0.00404   5.724 1.04e-08 ***
```

A comparison with linear models

- Linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

The coefficient β_1 represents the change of **E(Y)** if the covariate X1 increases by 1 unit.

- Logistic regression model

$$\log \frac{\Pr(Y=1)}{1-\Pr(Y=1)} = \log \frac{p}{1-p} = \beta_0 + \beta_1 X1 + \beta_2 X_2$$

The probability versus the odds

- There are two ways to express the "chance" of observing Y=1
  - The **probability** parameter $p = \Pr(Y = 1)$, which is bounded by 0 and 1.
  - The **odds of success** (Y=1)

$$\text{odds} = \frac{p}{1-p}$$

which is not bounded and can take any value.

$$\log \frac{\Pr(Y=1)}{1-\Pr(Y=1)} = \log \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

The interpretation of β_1

$$\log(\text{odds}) = \log \frac{p}{1-p} = \beta_0 + \beta_1 X1 + \beta_2 X_2$$

- β_1 represents the change of the **log odds** when X1 increases by 1 unite.

On the odds scale

$$\text{odds} = \frac{p}{1-p} = e^{\beta_0} e^{\beta_1 X1} e^{\beta_2 X_2}$$

- e^{β_1} represent the change of the odds when X1 increases by 1 unit.

$$\frac{\text{odds(after)}}{\text{odds(before)}} = e^{\beta_1}$$

An example

```
> beta.lmod<-coef(lmod)
> round(beta.lmod,3)
(Intercept)      height        cigs
     -4.502       0.025       0.023
```

- Smoke 1 more cigarette per day, $\frac{\text{odds(after)}}{\text{odds(before)}} = e^{0.023} = 1.023$. The odds of developing heart disease increase by 2.3%.

- Smoke 20 more cigarettes per day $\frac{\text{odds(after)}}{\text{odds(before)}} = e^{20\times 0.023} = 1.584$. The odds of developing heart disease increase by 58.4%.

Keep these in mind

- If we increase the covariate X by x units,
  - [Linear regression models] βx reflects the **<u>difference</u>** between E(Y)'s after and before the increase.
  - [Logistic regression models] $e^{\beta x}$ reflects the **<u>ratio</u>** between the odds after and before the increase.

How to estimate the coefficients?

```
glm(formula = chd ~ height + cigs, family = binomial,
data = wcgs)


Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.50161    1.84186  -2.444   0.0145 *
height       0.02521    0.02633   0.957   0.3383
cigs         0.02313    0.00404   5.724 1.04e-08 ***
```

Principle 4: Maximum likelihood

- The likelihood function

$$L(\beta_0, \beta_1, \beta_2) = \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i} \text{ where } p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

- The log-likelihood function

$$\ell(\beta_0, \beta_1, \beta_2) = \log(L(\beta_0, \beta_1, \beta_2)) = \sum_{i=1}^{N} [y_i \eta_i - \log(1+e^{\eta_i})]$$

$$\text{where } \eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

Model inference

- Which variables are "significant"? Do we need to include "height" in the model?
  - Model-L: chd ~ height + cigs Likelihood function L_L
  - Model-S: chd ~ cigs Likelihood function L_S

- If the smaller model is right,

$$2 \log \frac{L_L}{L_S} = D_L - D_S \sim \chi^2$$

Please try …

```
### Compare the models with and without “height”
lmod.s<-glm(chd~cigs,family=binomial,wcgs)
anova(lmod.s,lmod,test="Chi")

### Drop one covariate each time
drop1(lmod,test="Chi")
```

```
> anova(lmod.s,lmod,test="Chi")
Analysis of Deviance Table

Model 1: chd ~ cigs
Model 2: chd ~ height + cigs
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      3152       1750
2      3151       1749  1  0.92025   0.3374

> drop1(lmod,test="Chi")
Single term deletions

Model:
chd ~ height + cigs
       Df Deviance    AIC     LRT Pr(>Chi)
<none>      1749.0 1755.0
height  1   1750.0 1754.0  0.9202   0.3374
cigs    1   1780.1 1784.1 31.0695 2.49e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Confidence intervals

```
> confint(lmod)
                    2.5 %       97.5 %
(Intercept) -8.13475465 -0.91297018
height      -0.02619902  0.07702835
cigs         0.01514949  0.03100534
```