GLM with binary response

Sami Cheong

1/25/2024

Heart disease example

What might affect the chance of getting heart disease? One of the earliest studies addressing this issue started in 1960 and used 3154 healthy men, aged from 39 to 59, from the San Francisco area. At the start of the study, all were free of heart disease. Eight and a half years later, the study recorded whether these men now suffered from heart disease along with many other variables that might be related to the chance of developing this disease.

We load a subset of this data (cigarette use, height, and heart disease) from the Western Collaborative Group Study described in Rosenman et al. (1975) and focus on just three of the variables in the dataset:

```
data(wcgs, package="faraway")
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
```

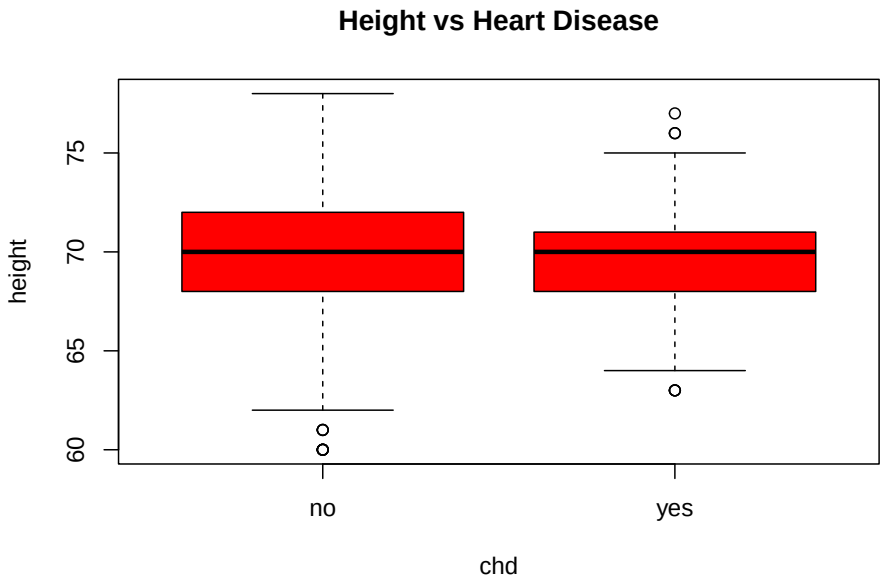
The variable `chd` indicates whether or not the person suffered from coronary heart disease, let's transform that to a binary variable

```
wcgs$y <- ifelse(wcgs$chd == 'yes',1,0)
```

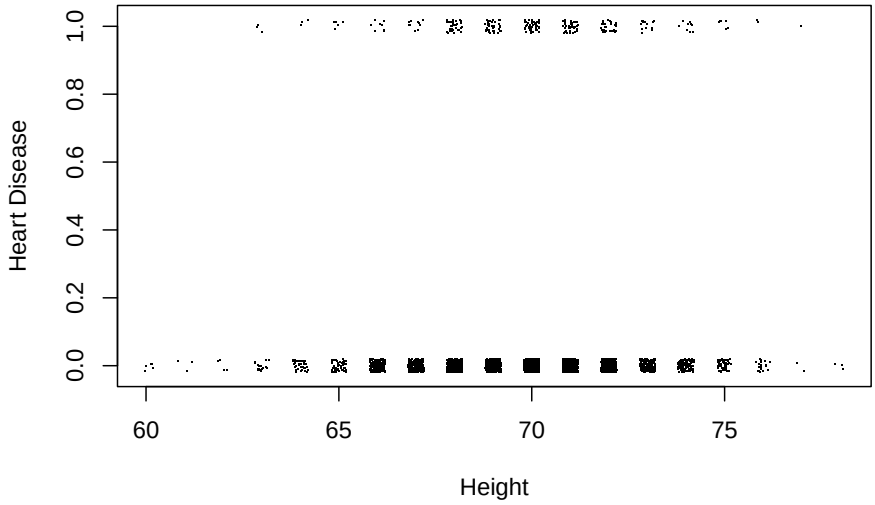
Exploratory data analysis:

Let's visualize the response and predictor variables in different dimensions to get a sense of how we should model it:

```
## boxplot view:
plot(height ~ chd, wcgs, col = 'red',
     main = 'Height vs Heart Disease' )
```

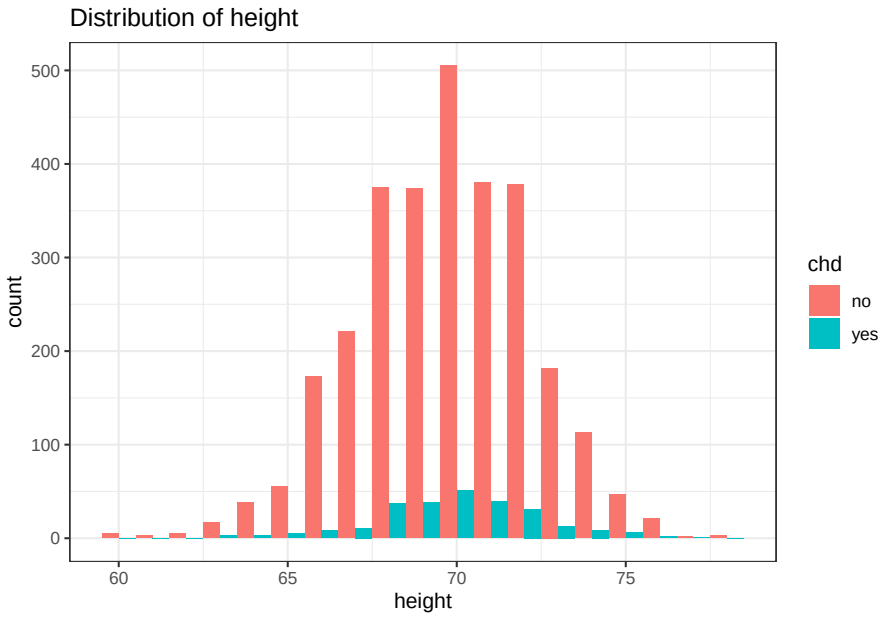



```
## jitter plot view
plot(jitter(y,0.1)~jitter(height),wcgs,
     xlab="Height", ylab="Heart Disease", pch=".")
```



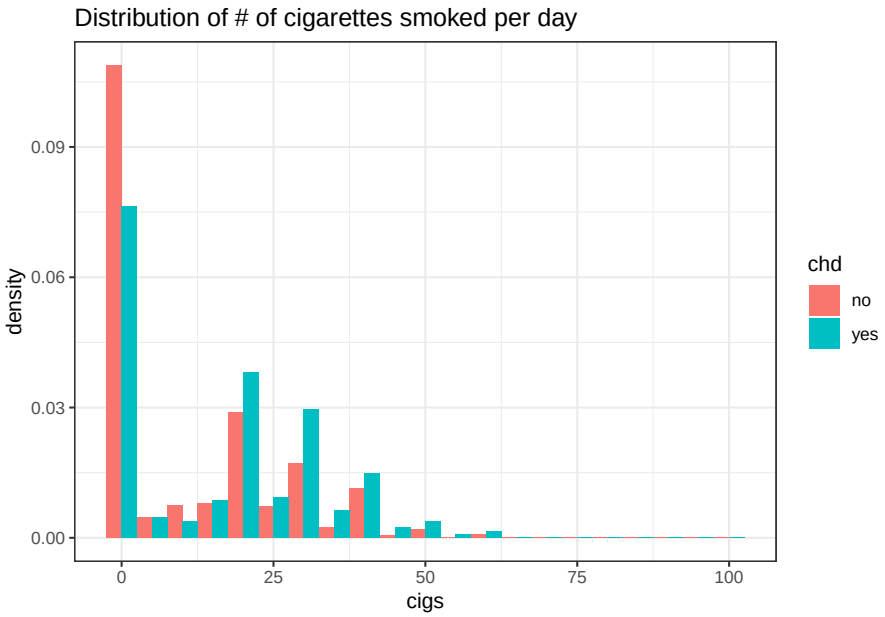
We can also use `ggplot` to explore the visualization, for example here are some plots of height and cigarette consumption information of the subjects separated by heart disease status:

```
ggplot(wcgs, aes(x=height, fill=chd)) +
  geom_histogram(position="dodge", binwidth=1)+
  theme_bw()+ggtitle('Distribution of height')
```



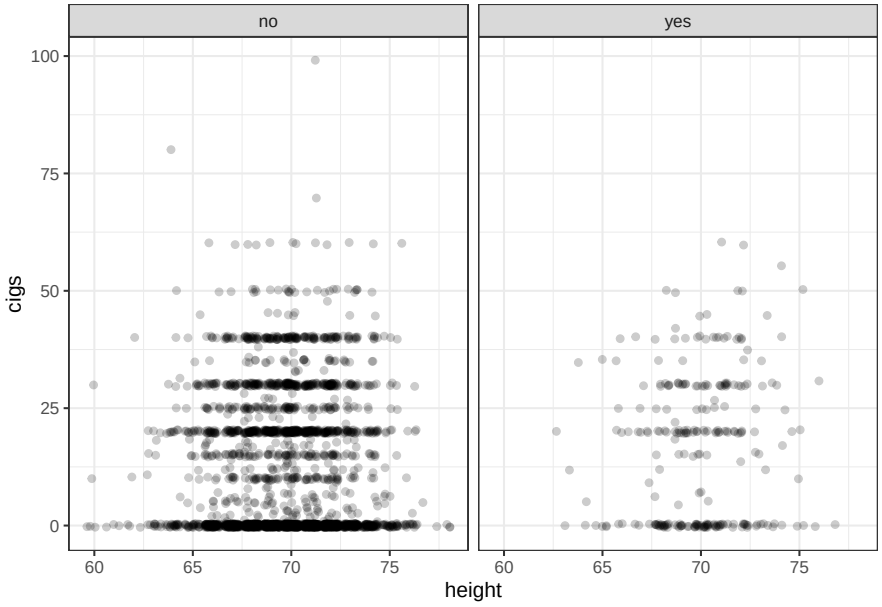

```
ggplot(wcgs, aes(x=cigs, fill=chd)) +
  geom_histogram(position="dodge", binwidth=5,
                 aes(y=..density..))+
  ggtitle('Distribution of # of cigarettes smoked per day')+theme_bw()
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



We can also separate the plots based on a chosen group using a `ggplot` feature called `facets`

```
ggplot(wcgs, aes(x=height,y=cigs))+
  geom_point(alpha=0.2, position=position_jitter())+
  facet_grid(~ chd)+theme_bw()
```



Analysis objective:

Our goal here is to predict the heart disease outcome for a given individual and also to explain the relationship between height, cigarette usage and heart disease.

Notice that for the same height and cigarette consumption, both outcomes can occur. It makes more sense then to model the response as a probability instead.

Model building :

Let y_i be a binary variable indicating presence or absence of heart disease (1, or 0). We would like to model it as the following :

$$p(\text{Heart disease for subject} i) = h(\beta_0 + \beta_1(\text{Height}) + \beta_2(\text{Cigarette use})$$

where $h = g^{-1}, g(\cdot)$ is the link function.

We can implement this in R as:

```
glmod <- glm(formula = chd ~ height + cigs,
             family = binomial(link = 'logit'),
             data = wcgs)
```

Understanding the model output:

Let's take a look at the summary of the model object `glmod`:

```
summary(glmod)
```

```
##
## Call:
## glm(formula = chd ~ height + cigs, family = binomial(link = "logit"),
##     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 ***
## ---
## 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
```

Recall the model formula is the following:

$$p(\text{Heart disease for subject}i) = h(\beta_0 + \beta_1(\text{Height}) + \beta_2(\text{Cigarette use})$$

since we chose the logit function as the link function, this can either be expressed as

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1(\text{Height}) + \beta_2(\text{Cigarette use})$$

Model coefficients : in here the β values correspond to the change of the log odds of the response. Notice the range of the log odds is $(-\infty, \infty)$ vs $[0, 1]$

```
beta <- coef(glmod)
print(beta)
```

```
## (Intercept)      height        cigs
## -4.50161397  0.02520779  0.02312740
```

For example, keeping everything else fixed, increase 1 cigarette smoked per day is associated with increasing the log odds of having heart disease by 0.023, or 2.3%.

Another way to interpret the model result is to look at the odds instead:

$$\text{odds of heart disease} = e^{\beta_0} e^{\beta_1 \text{Height}} e^{\beta_2 \text{Cigs}}$$

The model output is the follows:

```
print(exp(beta))
```

```
## (Intercept)      height        cigs
##  0.01109108  1.02552819  1.02339691
```

This tells us that a unit increase in , for example, cigarettes smoked per day, increases the odds of heart disease by a factor of e^{β_1} or 1.023, which is again about an 2.3% increase.

Null deviance: A low null deviance implies that the data can be modeled using only the intercept (without other predictors). If the null deviance is low, we should consider using fewer features for modeling the data.

Residual deviance: A low residual deviance implies that the model we have trained is appropriate.