Data-based Statistical Decision Model

Lecture 9 supplement - Predicting binary outcomes

Sungkyu Jung

Logistic regression

- A simple-minded understanding of logistic regression is to predict $y \in \{0, 1\}$ using $x_1, \ldots, x_p$ by a formula

$$y \sim \phi(x_1, \ldots, x_p) = \phi(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p) = \phi(\beta' X)$$

- Logistic regression starts with different model setup than linear regression: instead of modeling Y as a function of $X = (x_1, \ldots, x_p)$ directly, we model the probability that Y is equal to class 1, given X. First, abbreviate $p(X) = P(Y = 1|X)$. Then the logistic model is

$$p(X) = \frac{\exp(\beta' X)}{1 + \exp(\beta' X)}.$$

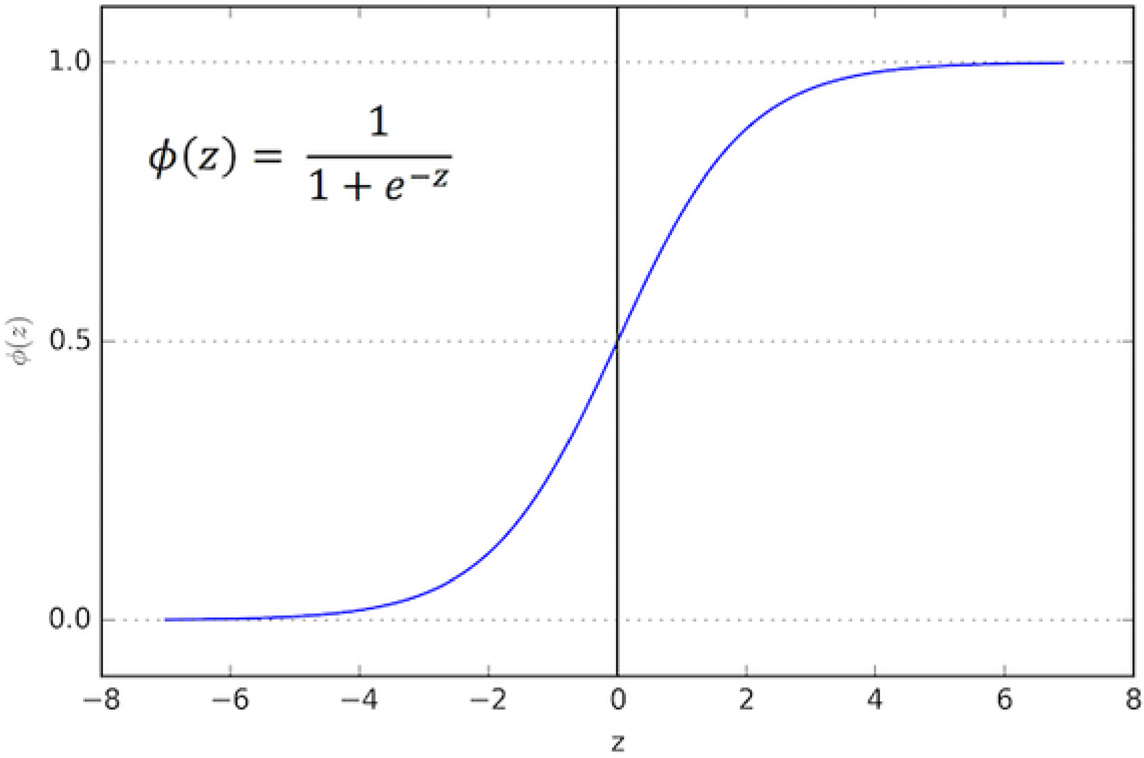
Logistic regression as a *generalized* linear regression

- "Linear": Explanatory variables to an intermediate predictor z:

$$(x_1, \ldots, x_p) \to z = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

- "Generalized": z to estimated probability p:

$$p = \phi(z)$$



Interpreting coefficients

- How can we interpret the role of the coefficients β?
- The logistic model is rearranged:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta' X$$

- The left-hand side $\log\left(\frac{p(X)}{1-p(X)}\right)$ is called *log-odds* of Class 1.
  - $p(X) = P(Y = 1 \mid X)$ = Probability of Class 1 (given X).
  - $p(X)/(1 - p(X))$ = *Odds* of Class 1 (given X)

Interpreting odds

Odds are an alternative scale to probability for representing chance

- As a way to express the payoffs for bets
- *Evens* bet: Winner gets paid an equal amount to that staked [Odds = 1]
- 3-1 *against* bet: pay $3 for every $1 [Odds = 1/3]
- 3-1 *on* bet: pay $1 for every $3 [Odds = 3]
- if the games are fair, if you win with probability p, then you would make in the long run

$$E(payout) = (1-p)(-1) + p(\frac{1}{\text{Odds}}) = 0$$

- That is, Odds = $\frac{p}{1-p}$.

A simple 1-d example

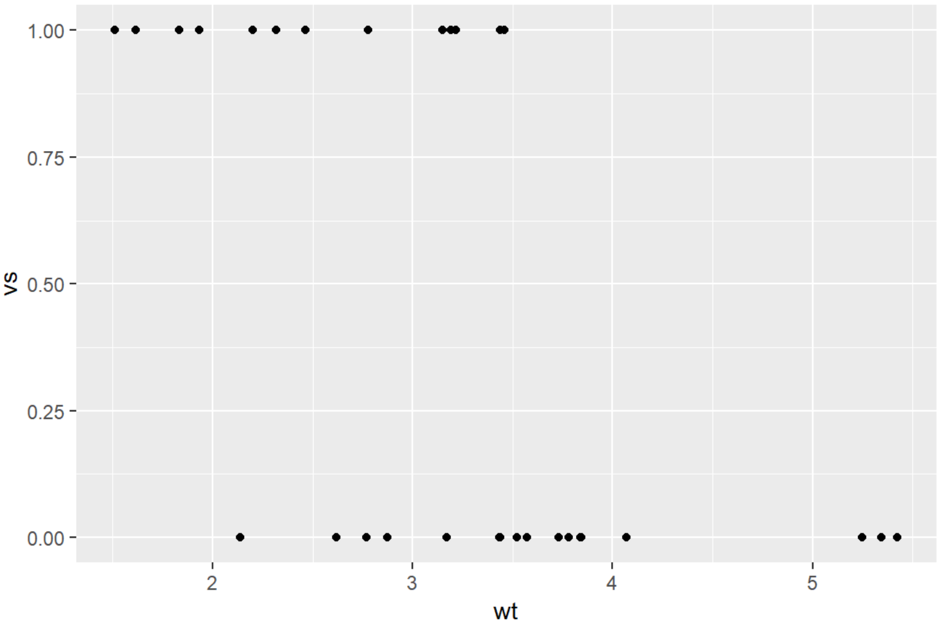
Motor Trend Car Road Tests

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). We will look at the following two variables:

- `wt` : Weight (1000 lbs)
- `vs` : Engine (0 = V-shaped, 1 = straight)

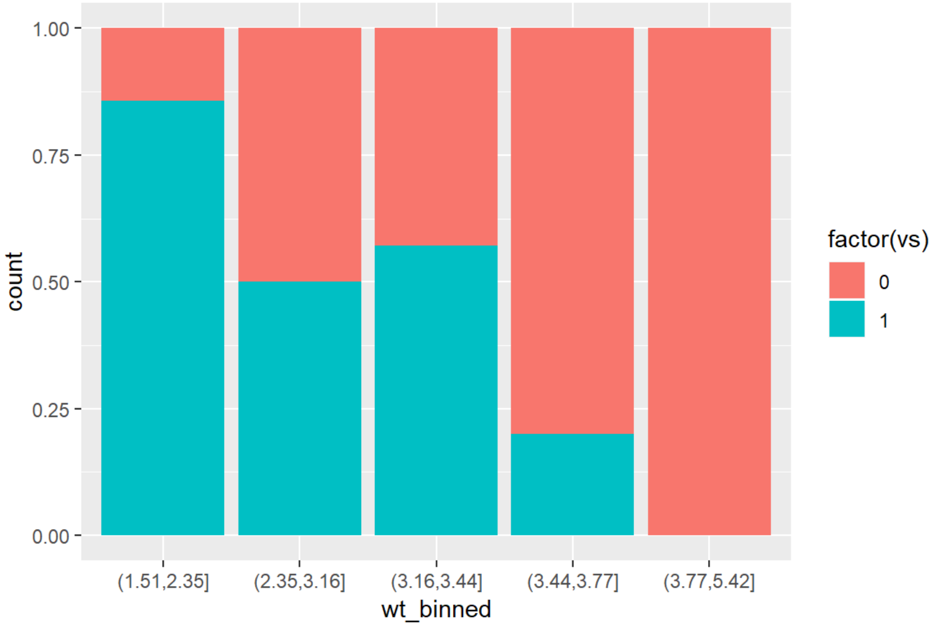
The data set is simple enough to visualize. What are we modeling here?

```
data(mtcars)
mtcars %>% ggplot(aes(x = wt, y = vs)) + geom_point()
```



The conditional probability is easier to see if we bin the observations.

```
wt_f <- unique(quantile(mtcars$wt, probs = seq(0,1, by = 0.2)))
wt_f[1] = wt_f-(1e-10)
mtcars %>% mutate(wt_binned = cut(wt, breaks = wt_f) ) %>% ggplot(aes(x = wt_binned, fill = factor(v
s))) + geom_bar(position = "fill")
```

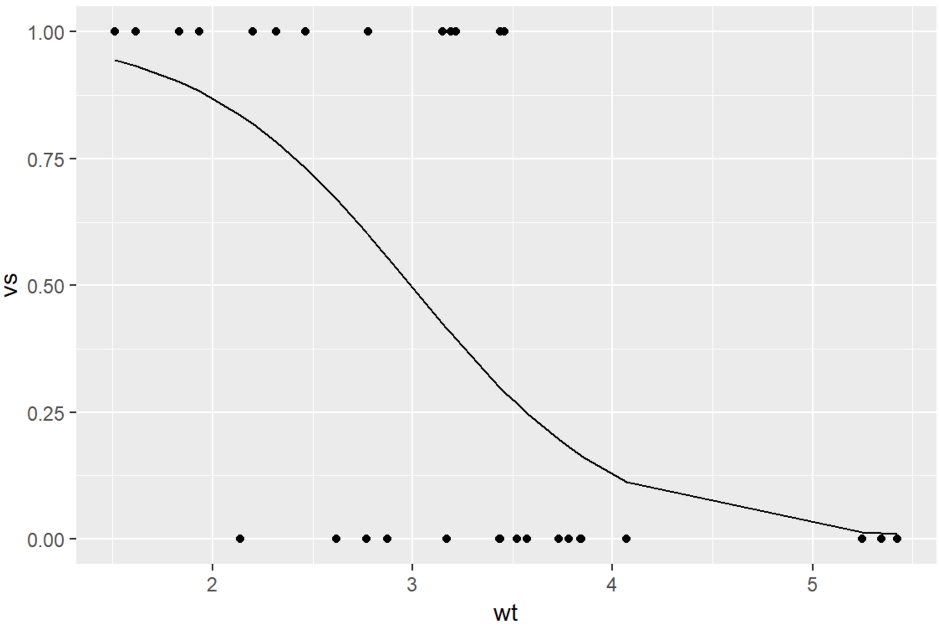


```
out <- glm(vs ~ wt, data = mtcars, family = "binomial")
summary(out)
```

```
##
## Call:
## glm(formula = vs ~ wt, family = "binomial", data = mtcars)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9003  -0.7641  -0.1559   0.7223   1.5736
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   5.7147     2.3014   2.483  0.01302 *
## wt           -1.9105     0.7279  -2.625  0.00867 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43.860  on 31  degrees of freedom
## Residual deviance: 31.367  on 30  degrees of freedom
## AIC: 35.367
##
## Number of Fisher Scoring iterations: 5
```

1. What is the model formula?
2. What are the odds of having a straight engine (`vs = 1`) if `wt` = 4000 lbs?
3. How does the odds change if the weight of the car increases by 1000 lbs?
4. Under which value of `wt` are the odds even?
5. Is there a simple explanation for the relation between β_1 and the conditional probability?

```
mtcars %>% ggplot(aes(x = wt, y = vs)) +
  geom_point() +
  geom_line(aes(x = wt, y = out$fitted.values)) +
  labs(main = "Data with fitted probability")
```



A bit more involved example

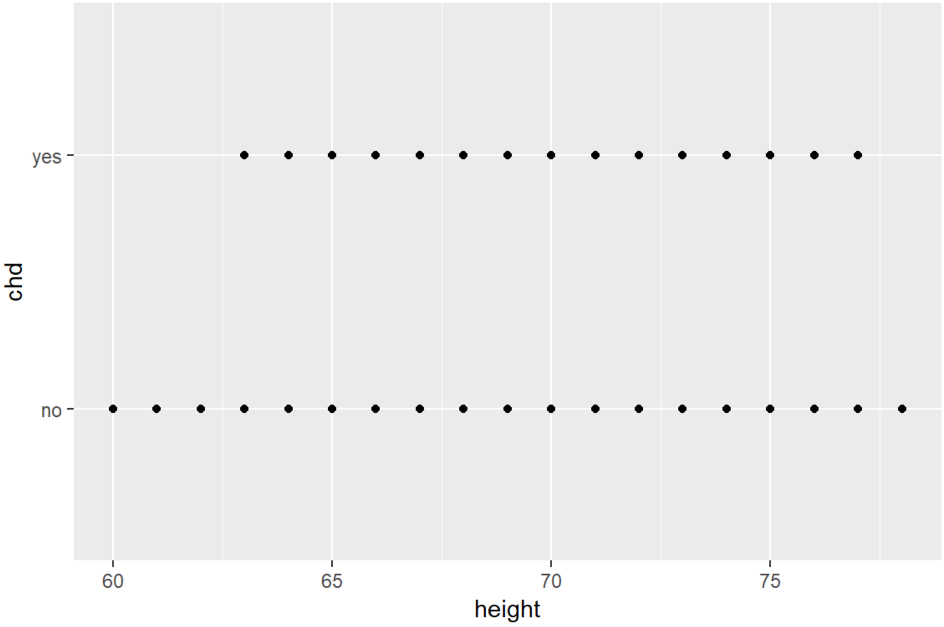
Western Collaborative Group Study

3154 healthy young men aged 39-59 from the San Francisco area were assessed for their personality type. All were free from coronary heart disease at the start of the research. Eight and a half years later change in this situation was recorded.

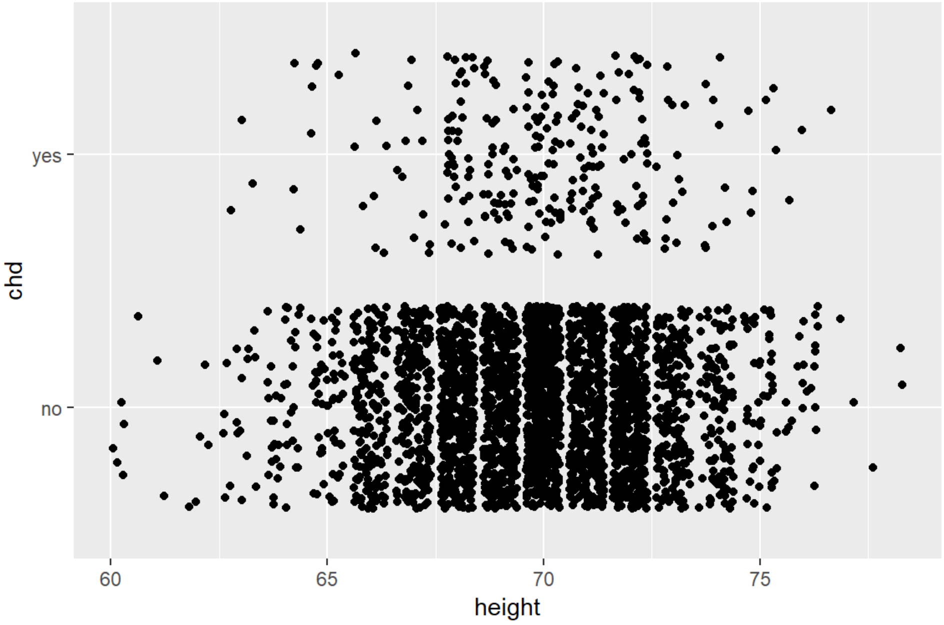
- `chd` : coronary heat disease developed is a factor with levels no yes
- `dibep` : behavior type a factor with levels A (Agressive) B (Passive)
- `height` : height in inches

Model `chd ~ height`

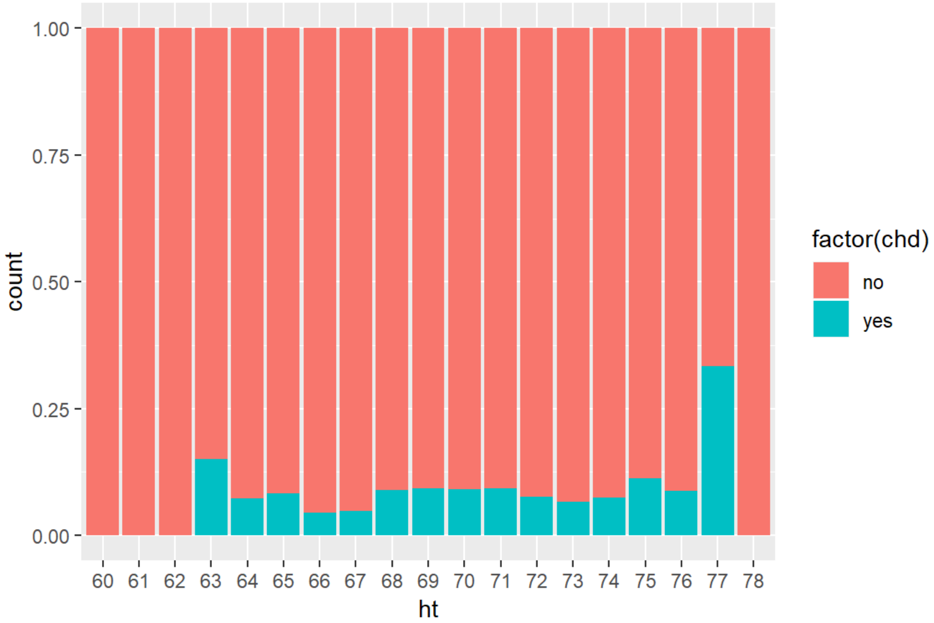
```
library(faraway)
data(wcgs)
wcgs %>% ggplot(aes(x = height, y = chd)) + geom_point() # What's going on here?
```



```
wcgs %>% ggplot(aes(x = height, y = chd)) + geom_jitter()
```



```
wcgs %>% mutate(ht = factor(height)) %>% ggplot(aes(x = ht, fill = factor(chd))) + geom_bar(position =
 "fill")
```



```
out <- glm(chd ~ height, data = wcgs, family = "binomial")
summary(out)
```

```
##
## Call:
## glm(formula = chd ~ height, family = "binomial", data = wcgs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4587  -0.4186  -0.4131  -0.4024   2.3157
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.33732    1.81231  -2.393   0.0167 *
## height       0.02742    0.02590   1.058   0.2899
## ---
## 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: 1780.1  on 3152  degrees of freedom
## AIC: 1784.1
##
## Number of Fisher Scoring iterations: 5
```

```
height_grd <- seq(min(wcgs$height),max(wcgs$height), 1)
height_pred <- predict(out, newdata=  data.frame(height = height_grd), type = "response")
tibble(x = height_grd, y = height_pred)
```

```
## # A tibble: 19 x 2
##        x      y
##    <dbl>  <dbl>
##  1    60 0.0634
##  2    61 0.0651
##  3    62 0.0668
##  4    63 0.0685
##  5    64 0.0703
##  6    65 0.0721
##  7    66 0.0739
##  8    67 0.0758
##  9    68 0.0778
## 10    69 0.0798
## 11    70 0.0818
## 12    71 0.0839
## 13    72 0.0860
## 14    73 0.0882
## 15    74 0.0904
## 16    75 0.0927
## 17    76 0.0950
## 18    77 0.0974
## 19    78 0.0998
```

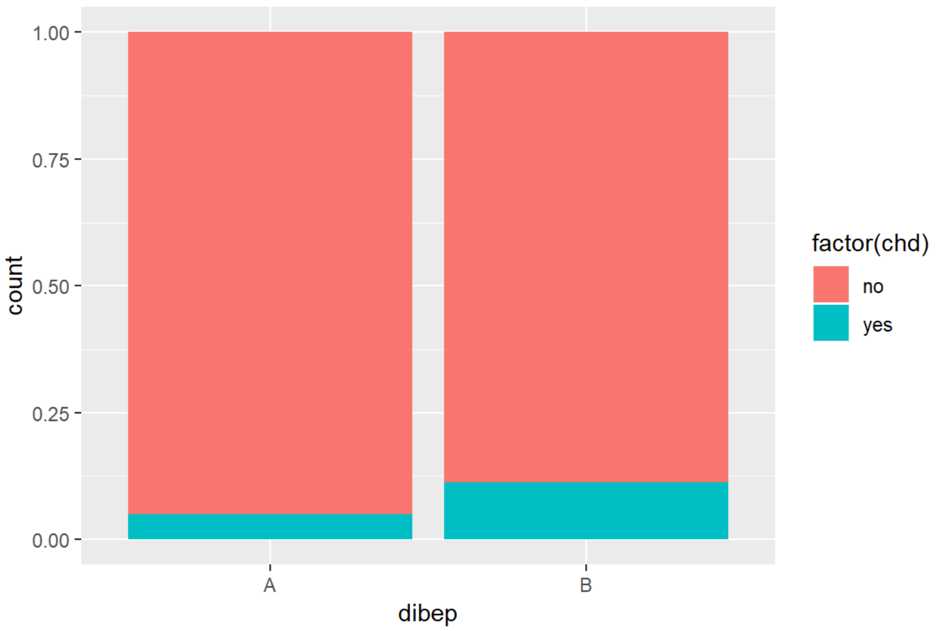
Model `chd ~ dibep`

Both variables are binary. Logistic regression becomes quite simple.

```
with(wcgs, table(chd,dibep))
```

```
##      dibep
## chd      A    B
##   no  1486 1411
##   yes   79  178
```

```
wcgs %>% ggplot(aes(x = dibep, fill = factor(chd))) + geom_bar(position = "fill")
```



```
out <- glm(chd ~ dibep, data = wcgs, family = "binomial")
summary(out)
```

```
## 
## Call:
## glm(formula = chd ~ dibep, family = "binomial", data = wcgs)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.4875  -0.4875  -0.3219  -0.3219   2.4438  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -2.9344     0.1155 -25.416  < 2e-16 ***
## dibepB        0.8641     0.1402   6.163 7.12e-10 ***
## ---
## 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: 1740.3  on 3152  degrees of freedom
## AIC: 1744.3
## 
## Number of Fisher Scoring iterations: 5
```

```
unique(out$fitted.values)
```

```
## [1] 0.11202014 0.05047923
```

1. What is the model?
2. Can you interpret β_1?
3. What is the probability of coronary heat disease when the behavior type is passive? Agressive?