Generalized Linear Models

Alexander McLain

Contents

1 Logistic regression example

```
library(tidyverse)
data_adult <-read.csv("https://raw.githubusercontent.com/guru99-edu/R-Programming/master/adult.csv")
glimpse(data_adult)
```

```
## Rows: 48,842
## Columns: 10
## $ x               <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ...
## $ age             <int> 25, 38, 28, 44, 18, 34, 29, 63, 24, 55, 65, 36, 26,...
## $ workclass       <chr> "Private", "Private", "Local-gov", "Private", "?", ...
## $ education       <chr> "11th", "HS-grad", "Assoc-acdm", "Some-college", "S...
## $ educational.num <int> 7, 9, 12, 10, 10, 6, 9, 15, 10, 4, 9, 13, 9, 9, 9, ...
## $ marital.status  <chr> "Never-married", "Married-civ-spouse", "Married-civ...
## $ race            <chr> "Black", "White", "White", "Black", "White", "White...
## $ gender          <chr> "Male", "Male", "Male", "Male", "Female", "Male", "...
## $ hours.per.week  <int> 40, 50, 40, 40, 30, 30, 40, 32, 40, 10, 40, 40, 39,...
## $ income          <chr> "<=50K", "<=50K", ">50K", ">50K", "<=50K", "<=50K",...
```

When we do logistic regression for one variable X_i, we assume that

$$P(Y_i = 1|X_i) = p_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}.$$

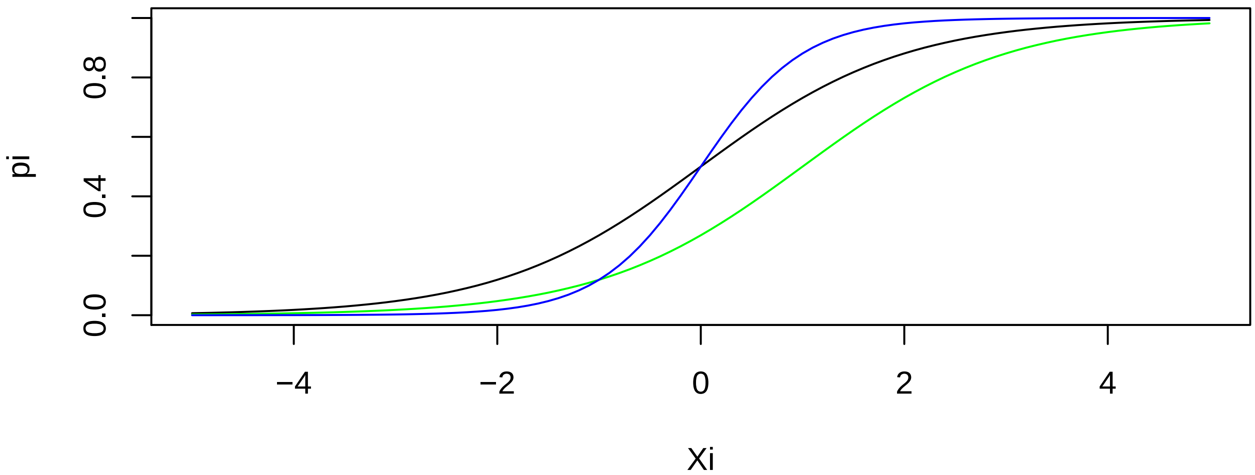
Let's say that $\beta_0 = 0$ and $\beta_1 = 1$, in this case the relationship between X_i and p_i looks like the following:

```
Xi <- seq(-5, 5, 0.1)
pi = exp( Xi)/( 1 + exp( Xi) )
plot( Xi, pi, type = "l")

#Let's see what changing $\beta_0$ and $\beta_1$ do:

pi = exp( -1 + Xi)/( 1 + exp( -1 + Xi) )
lines( Xi, pi, col = "green")

pi = exp( 2*Xi)/( 1 + exp( 2*Xi) )
lines( Xi, pi, col = "blue")
```




```r
table( data_adult$income)
```

<=50K	>50K
37155	11687

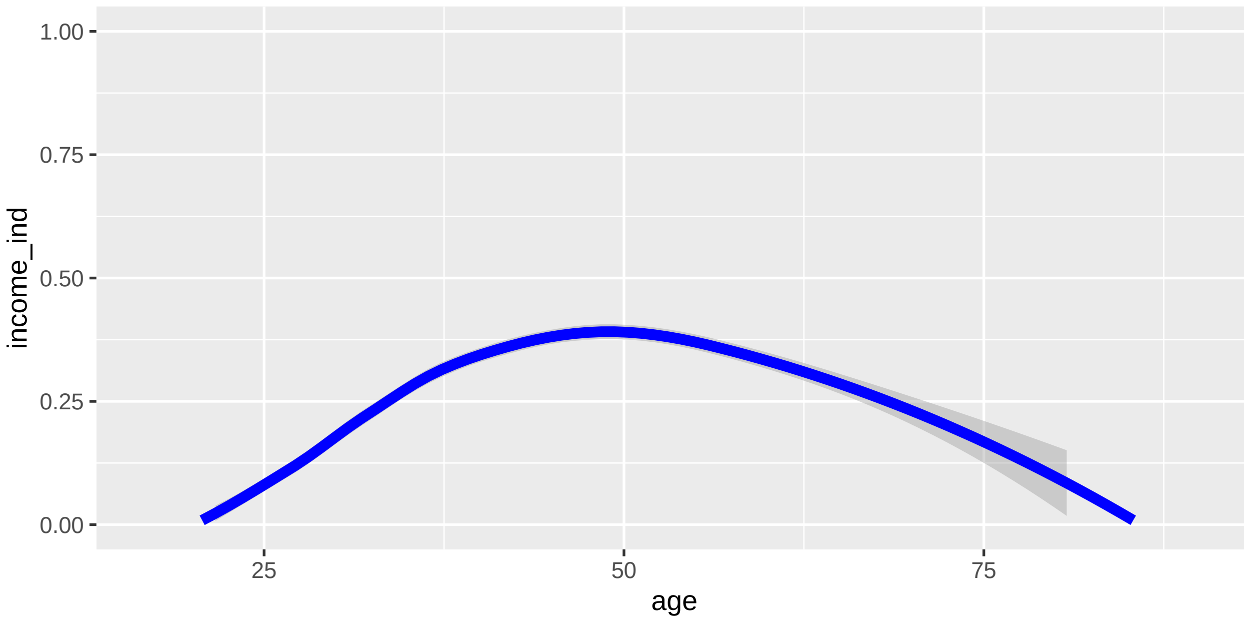
```r
data_adult <- data_adult %>% mutate( income_ind =
                                       ifelse(income == ">50K",1,0) )
data_adult_sm <- data_adult %>% filter(x < 10001)
table( data_adult_sm$income_ind)
```

0	1
7643	2357

```r
p <- ggplot(data_adult_sm, aes(x = age, y = income_ind) )
p + geom_smooth(aes(group = 1), method = "loess",
                              color = "blue", size = 2) +
  ylim(0,1)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 9 rows containing missing values (geom_smooth).
```



Let's try analyzing this with glm

```
mean(data_adult$age)
```

```
## [1] 38.64359
```

```
sd(data_adult$age)
```

```
## [1] 13.71051
```

```
data_adult <- data_adult %>% mutate( age = scale(age),
                                     age2 = scale(age)^2)
formula <- income_ind ~ age + age2
log_reg <- glm( formula, family = binomial( link = "logit"), data = data_adult)
summary(log_reg)
```

```
##
## Call:
## glm(formula = formula, family = binomial(link = "logit"), data = data_adult)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0458  -0.8746  -0.4933  -0.1759   3.3872
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.79231    0.01379  -57.47   <2e-16 ***
## age          1.11010    0.01793   61.93   <2e-16 ***
## age2        -0.64852    0.01332  -48.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 53751  on 48841  degrees of freedom
## Residual deviance: 47775  on 48839  degrees of freedom
```

```
## AIC: 47781
##
## Number of Fisher Scoring iterations: 5
```

```
Xi <- seq(-5, 5, 0.1)
pi = exp(-0.79231 + 1.11010*Xi - 0.64852*Xi^2)/( 1 + exp( -0.79231 + 1.11010*Xi - 0.64852*Xi^2) )
age_vec <- Xi*13.71051 + 38.64359
plot( age_vec, pi, type = "l", xlim = c( 20, 80), ylim = c( 0, 1))
```