W271 Summer 2022 Lecture Video Question Solutions Week 1

Contents

Week 2 Discrete Response Model Part 2

2.2 Introduction to Binary Response Models and the Linear Probability Model

Q: Explain why the following is true: when y is a binary variable taking on values 0 and 1, we have $E(y|x) = P(y|x)$.

Solution: This is very similar to our calculation of the mean of the binomial probability distribution in week 1.

$E(y|x) = 1 * P(y = 1|x) + 0 * P(y = 0|x) = P(y = 1|x)$

Q: Derive the conditional variance of the LPM: $Var(y|x) = p(\mathbf{x})(1 - p(\mathbf{x}))$.

Solution: This is very similar to our calculation of the variance of the binomial probability distribution in week 1.

$Var(y|x) = E(y^2|x) - E(y|x)^2 = [1^2 * P(y = 1|x) + 0^2 * P(y = 0|x)] - P(y = 1|x)^2 = P(y = 1|x) - P(y = 1|x)^2 = P(y = 1|x)(1 - P(y = 1|x))$

We can denote $P(y = 1|x) = P(\mathbf{x})$ since the probability of success is a linear function of the $\mathbf{x}$ values.

2.3 The Logit Transformation and the Logistic Curve

Q: Derive the logistic function from the log (odd ratio) form.

Solution: In logistic regression we model the log odds of success as:

$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 x_{i1} ... + \beta_p x_{ip}$.

Therefore:

$\frac{\pi_i}{1-\pi_i} = e^{\beta_0 + \beta_1 x_{i1} ... + \beta_p x_{ip}}$

$$\pi_i = e^{\beta_0+\beta_1 x_{i1}\ldots+\beta_p x_{ip}}(1-\pi_i)$$

$$\pi_i + \pi_i e^{\beta_0+\beta_1 x_{i1}\ldots+\beta_p x_{ip}} = e^{\beta_0+\beta_1 x_{i1}\ldots+\beta_p x_{ip}}$$

$$\pi_i(1+e^{\beta_0+\beta_1 x_{i1}\ldots+\beta_p x_{ip}}) = e^{\beta_0+\beta_1 x_{i1}\ldots+\beta_p x_{ip}}$$

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_{i1}\ldots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}\ldots+\beta_p x_{ip}}}$$

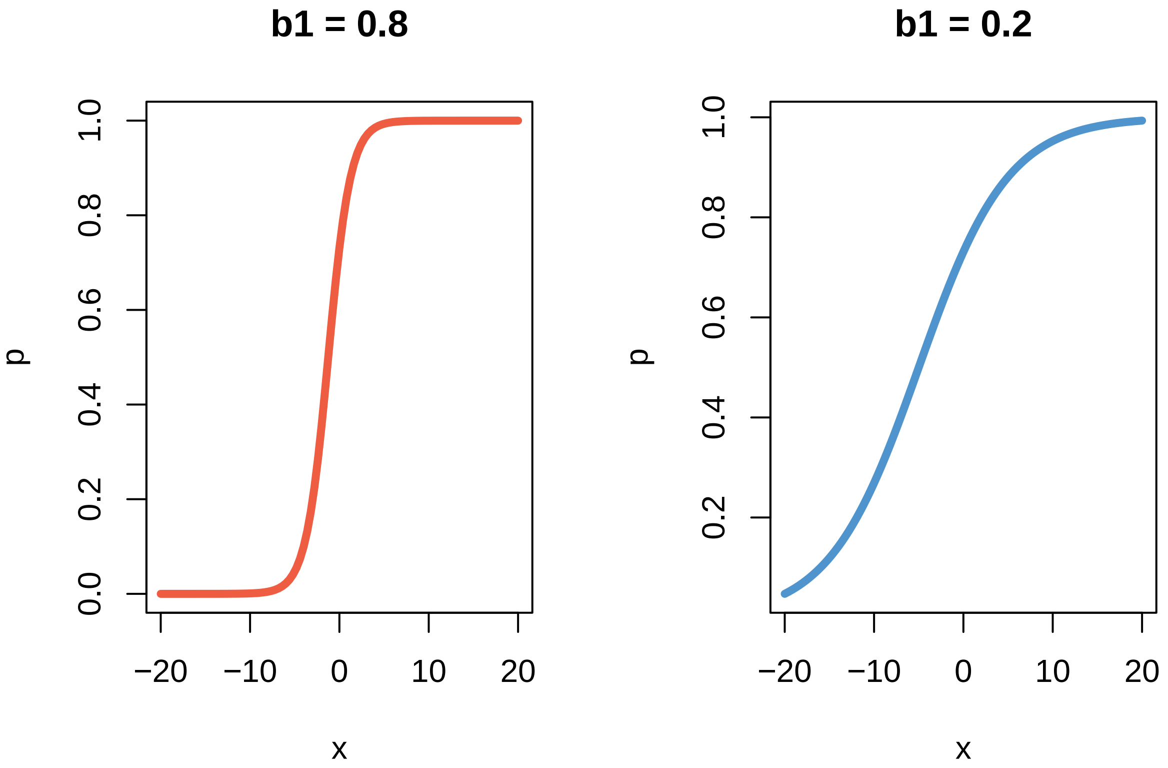
Q: Recreate the plots, assuming there is only one explanatory variable where $b_0 = 1$ and $b_1 = 0.8$ and $b_0 = 1$ and $b_1 = 0.2$.

Solution: We have $log(\frac{\pi_i}{1-\pi_i}) = b_0 + b_1 x_i$.

```
par(mfrow=c(1,2))

b0 <- 1; b1 <- 0.8
curve(expr = exp(b0 + b1*x)/(1 + exp(b0 + b1*x)), xlim = c(-20, 20), lwd = 4,
      xlab = "x", ylab = "p", col = "tomato2",
      main = paste("b1 = ", b1, sep = ""))

b0 <- 1; b1 <- 0.2
curve(expr = exp(b0 + b1*x)/(1 + exp(b0 + b1*x)), xlim = c(-20, 20), lwd = 4,
      xlab = "x", ylab = "p", col = "steelblue3",
      main = paste("b1 = ", b1, sep = ""))
```


b1 = 0.8

b1 = 0.2

Note how because $b_1 > 0$ both plots show a positive relationship between x and π. But when $b1 = 0.8$ the graph i flatter for a larger range of values of x and steeper for x close to zero compared to $b1 = 0.2$ where the slope is flatter and more gradual. This is because when x coefficients are larger in logistic regression, the impact to the probability is larger as $\frac{d\pi}{dx} = b_1\pi(1-\pi)$, which is increasing in b_1.

2.6 An Example

Q: Based on the model provided, answer the following questions: compute the predicted probability when change=0.5 and distance=50 and using this estimated model plot the estimated probability (of success) curve.

```
# data can be downloaded from https://www.chrisbilder.com/categorical/programs_and_data.html
df <- read.csv("~/Documents/Berkeley W271/Week 2 Discrete Response Model Part 2/Placekick.csv",
               stringsAsFactors = F)

glm.model <- glm(formula = good ~ change + distance, family = binomial(link = logit),
                 data = df)

summary(glm.model)
```

Solution:

```
## 
## Call:
## glm(formula = good ~ change + distance, family = binomial(link = logit), 
##     data = df)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.7061   0.2282   0.2282   0.3750   1.5649  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  5.893181   0.333184  17.687   <2e-16 ***
## change      -0.447783   0.193673  -2.312   0.0208 *  
## distance    -0.112889   0.008444 -13.370   <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: 1013.4  on 1424  degrees of freedom
## Residual deviance:  770.5  on 1422  degrees of freedom
## AIC: 776.5
## 
## Number of Fisher Scoring iterations: 6
```

(1) Compute the predicted probability when change=0.5 and distance=50.

```
new.data <- data.frame("change" = 0.5, "distance" = 50)

pred.prob <- predict(glm.model, newdata = new.data, type = "response")
```

```r
print(paste("P(good | change = 0.5, distance = 50) = ", round(pred.prob, 2), sep = ""))
```

```
## [1] "P(good | change = 0.5, distance = 50) = 0.51"
```

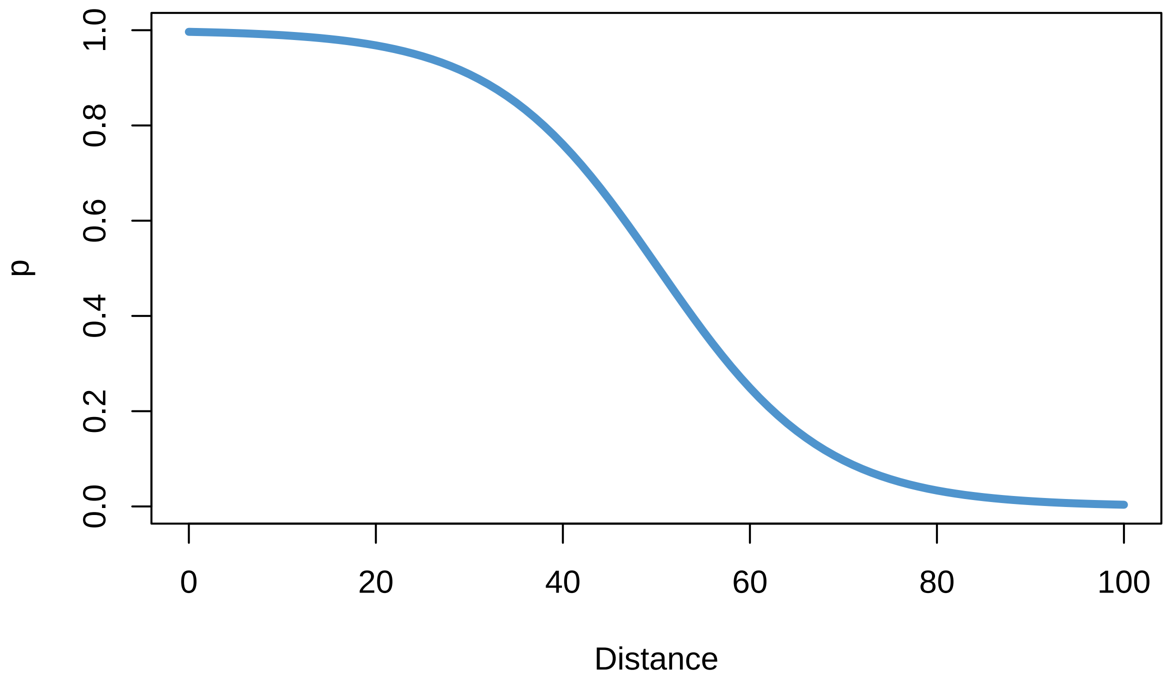
(2) Plot the estimated probability (of success) curve

We will fix change=0.5 for simplicity and plot the probability as a function of distance.

```r
# distance is in yards and a football field ranges from 0 to 100 yards
new.data <- data.frame("change" = 0.5, "distance" = 0:100)

pred.prob <- predict(glm.model, newdata = new.data, type = "response")

par(mfrow=c(1,1))
plot(x = new.data$distance, y = pred.prob, lwd = 4, col = "steelblue3",
     xlab = "Distance", ylab = "p", type = "l")
```



2.10 Odds Ratios

Q: Verify that the odds ratio form in the logistic regression model.

Solution: Remember that the odds ratio $OR = \frac{P(y=1|x_r=x+c)/P(y=0|x_r=x+c)}{P(y=1|x_r=x)/P(y=0|x_r=x)}$.

$log(\frac{P(y=1|x_r=x+c)}{P(y=0|x_r=x+c)}) = \beta_0 + \beta_1 * 0 + ... + \beta_r * (x+c) + .. + \beta_p * 0 = \beta_r(x+c)$

$\frac{P(y=1|x_r=x+c)}{P(y=0|x_r=x+c)} = e^{\beta_r(x+c)}$

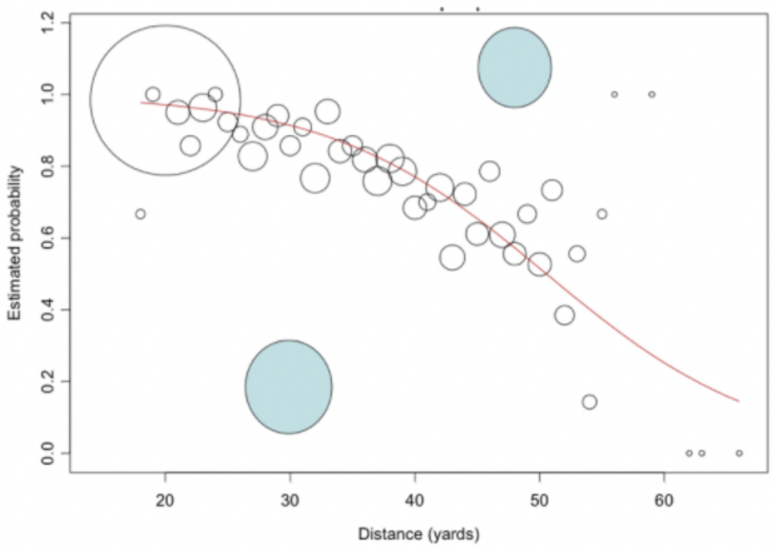
$log(\frac{P(y=1|x_r=x)}{P(y=0|x_r=x)}) = \beta_0 + \beta_1 * 0 + ... + \beta_r * x + .. + \beta_p * 0 = \beta_r x$

$\frac{P(y=1|x_r=x)}{P(y=0|x_r=x+c)} = e^{\beta_r x}$

$OR = \frac{e^{\beta_r(x+c)}}{e^{\beta_r x}} = e^{\beta_r(x+c) - \beta_r x} = e^{\beta_r c}$

2.12 Visual Assessment of the Logistic Regression Model

Q: What do you think about the fit of the model in the provided plot?



Solution:

The above plot shows the fitted probabilities from the glm model as a function of distance in red. We see various sized circles based on the frequency of observations at that distance.

Generally, the model fit looks pretty poor because there are huge clusters of observations at 30 yards and 50 yards that do not lie on the red curve, meaning they exhibit unusual chances of making the kick (light blue circles). It appears that the simple linear relationship assumed by the glm model is not fitting the data well or there are other factors for those kicks that we are accounting for.