Logistic Regression

(A Subtitle Would Go Here if you wish)

Honglang Wang

Department of Mathematical Sciences, IUPUI

- Up to now, we've used statistical methods to investigate the relationship of two variables when
  - the response variable is **quantitative** and the explanatory variable is **categorical** (comparing two means, ANOVA);
  - the response variable is **categorical** and the explanatory variable is **categorical** (comparing two proportions, contingency table);
  - the response variable is **quantitative** and the explanatory variable is **quantitative** (linear regression).
- *What if the response variable is **categorical** and the explanatory variable is **quantitative** ?*

- For example, a credit card company is concern about whether their card members would pay a bill on time or not. They may use the size of bill, annual income, and other information of members to predict the probability that they would pay the bill on time.
- In this unit, we are going to study the logistic regression model to deal with the situation when the explanatory variable is quantitative and the response variable variable is categorical (specifically, the response variable is binary "yes/no").

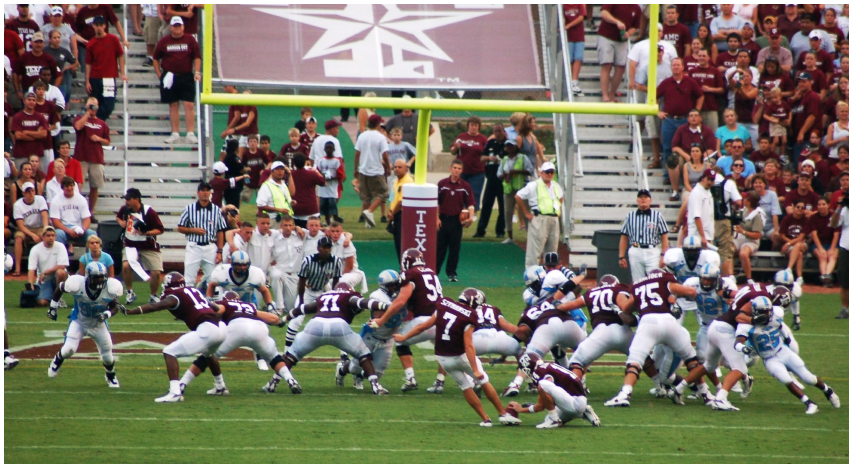
Logistic Regression

- Logistic regression model:

1. Use logistic regression to model the binary data given some quantitative explanatory variable
2. Estimate model parameters
3. Inference

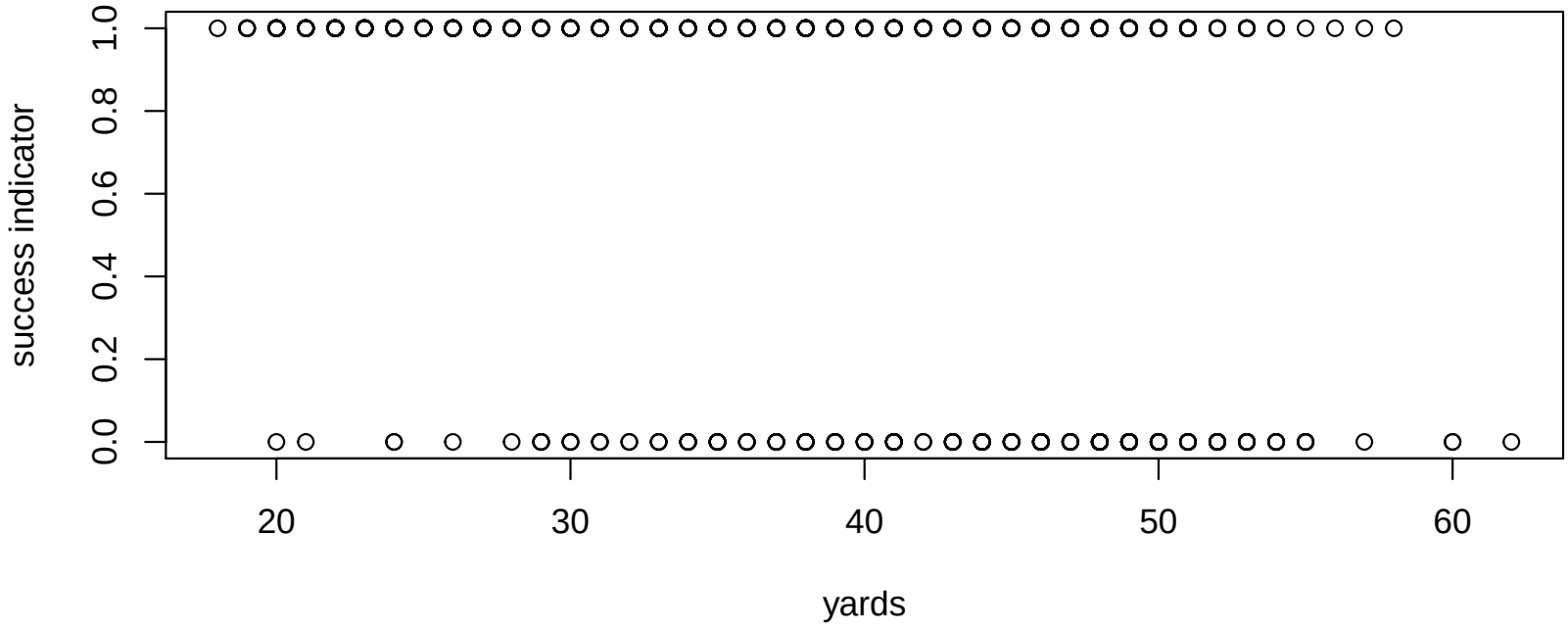
Logistic Regression

Example 9.1 The field goal is a critical scoring play in the National Football League (NFL). The attempt distance is the primary factor determining success. The data collected here includes all NFL field goal attempts in 2003. For each attempt, the attempt distance x (yards) and the success indicator y were recorded, where $y = 1$ represents success and $y = 0$ represents failure. Overall we have 948 pairs of data (x_i, y_i). The problem we are interested in here is how the probability of success would be affected by the attempt distance.



Logistic Regression

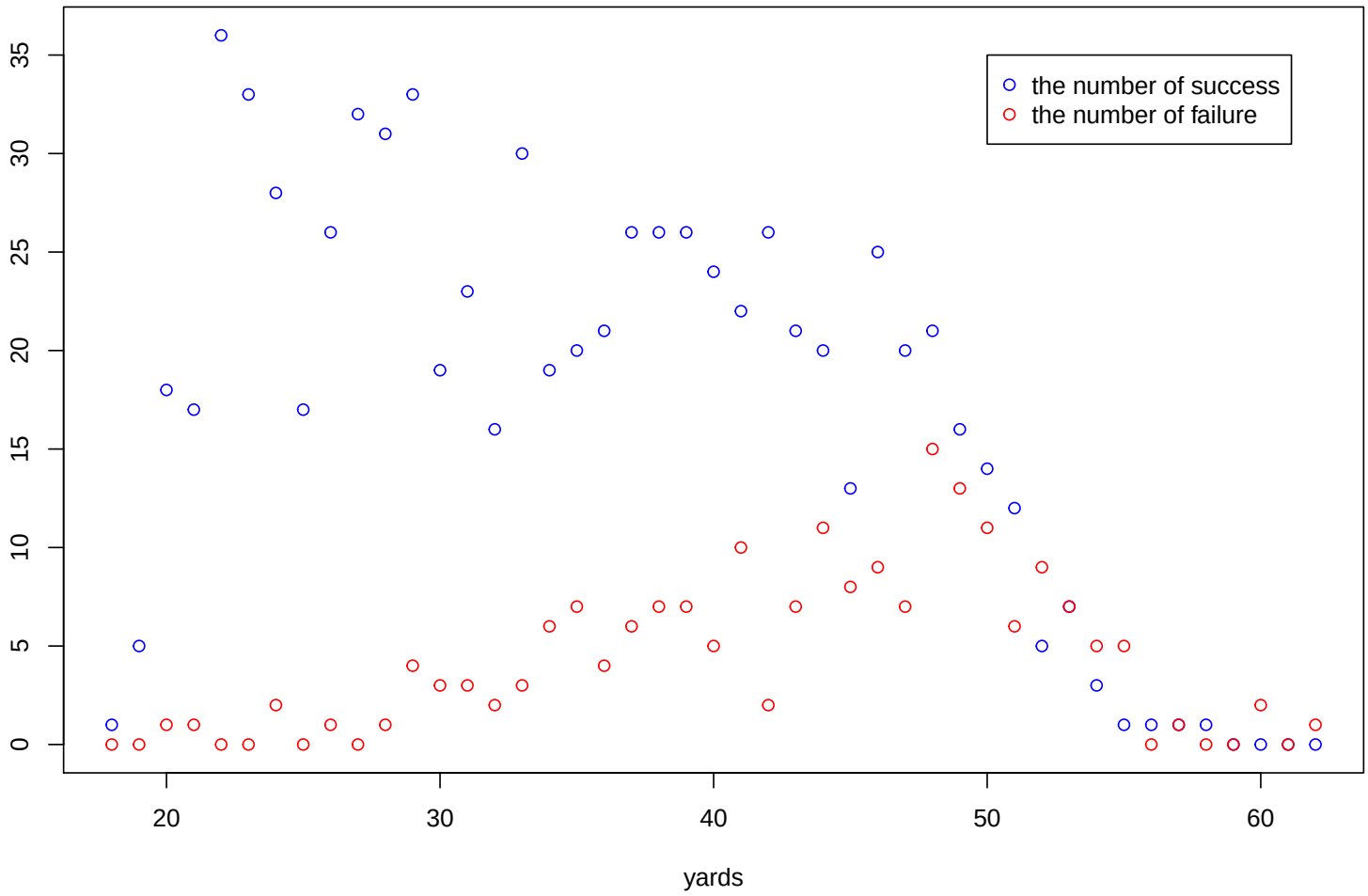
1. **Exploratory data analysis (abbr. EDA)**



Q1: Scatterplot? Good or not?
Q2: Simple linear model ($y = \beta_0 + \beta_1 x + \epsilon$)?

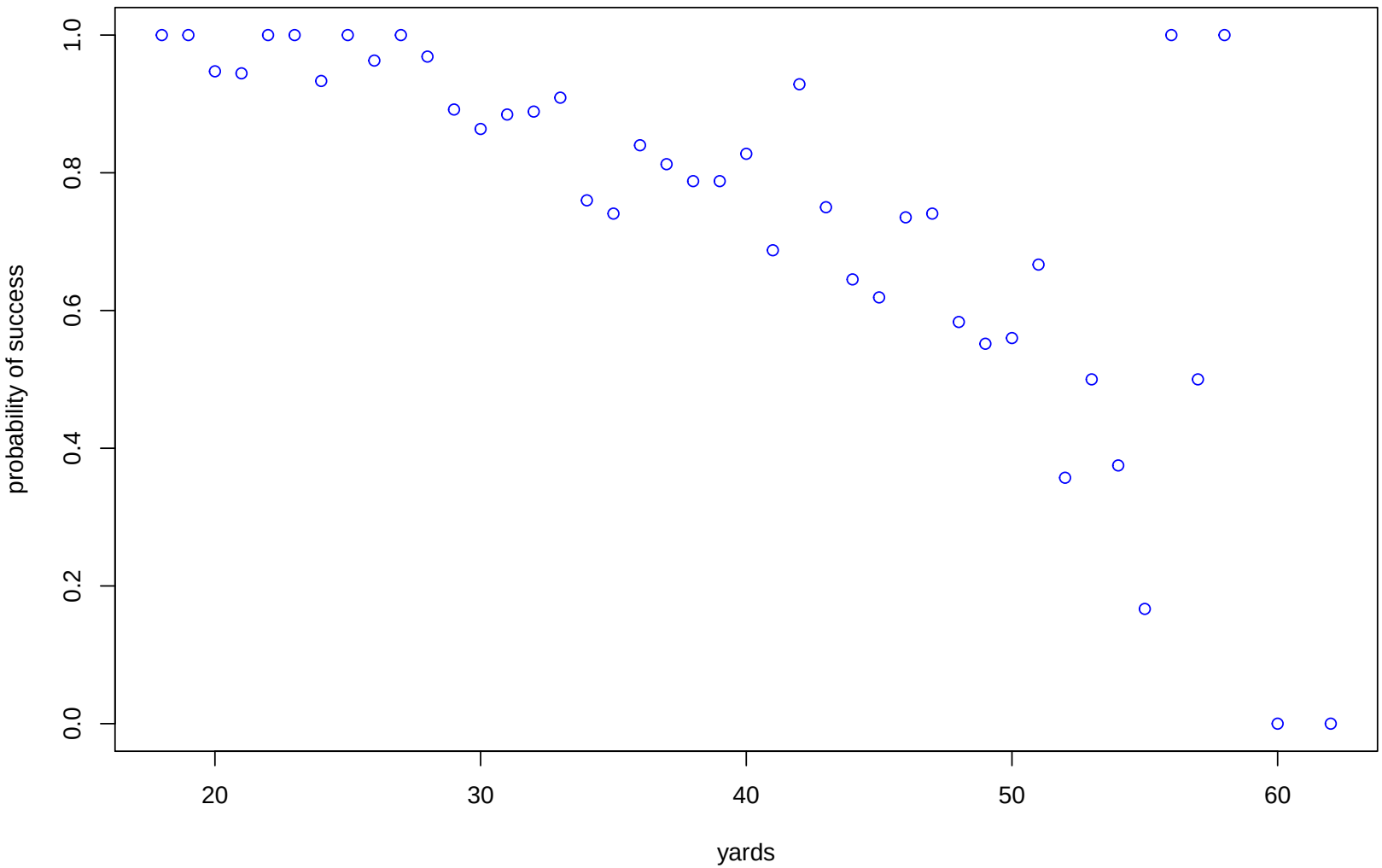
Logistic Regression

- Given attempt distance, count the number of success and the number of failure.



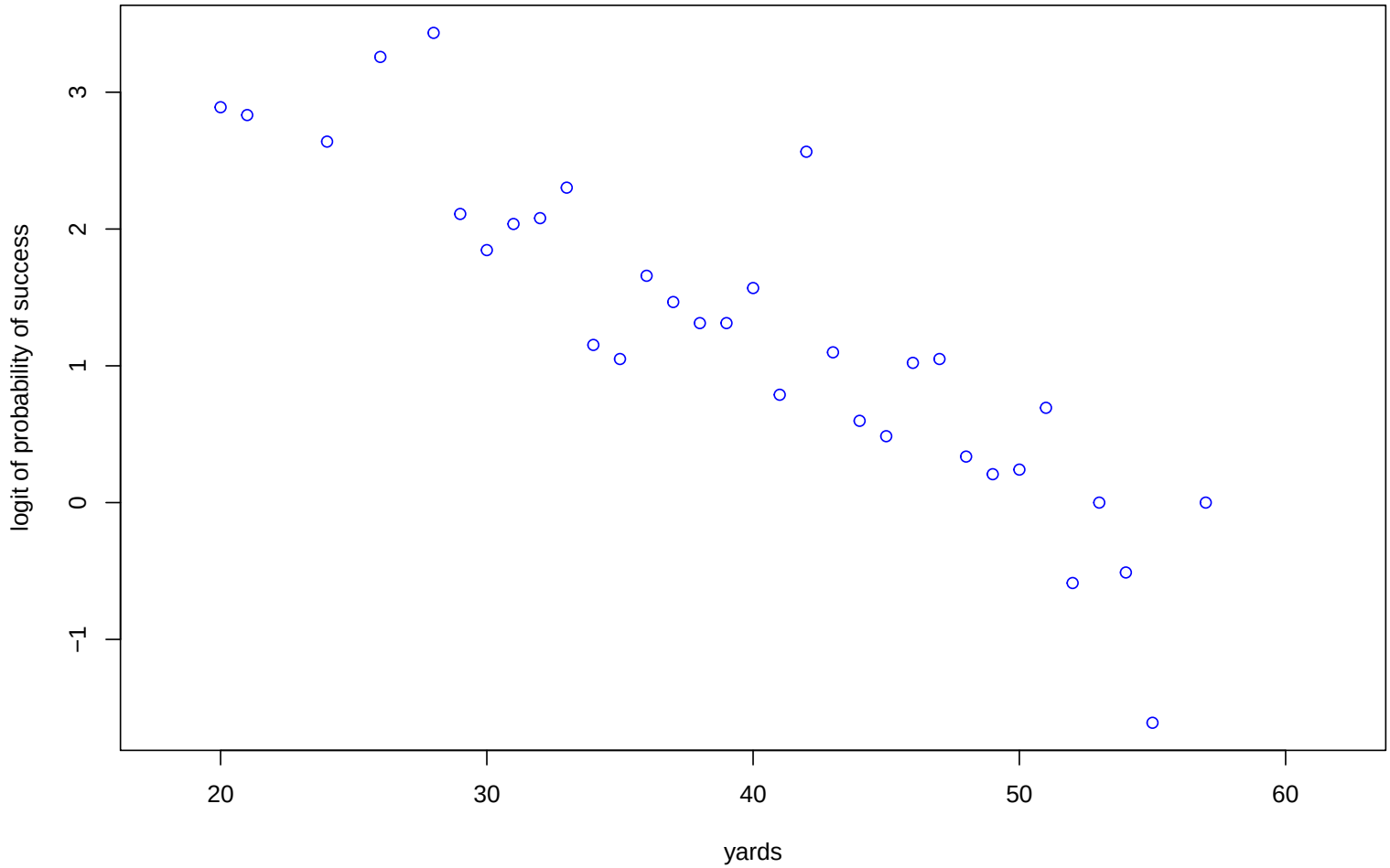
Logistic Regression

- Given attempt distance, calculate the probability of success. *Q: Linear pattern?*



Logistic Regression

- Perform logit transform on the probability of success $logit(\pi) = \log(\frac{\pi}{1-\pi})$. *Q: Linear pattern?*



Logistic Regression

2. **Simple logistic regression**

- Based on the previous EDA, we can see that the probability of success depends on the attempt distance of field goal.
- Given x, assume that the probability of success is $\pi(x)$. Then the corresponding success indicator y is distributed as

y	1	0
p(y)	$\pi(x)$	$1-\pi(x)$

Actually, y follows binomial distribution with $n = 1$ and $\pi = \pi(x)$.

Logistic Regression

- We found that $\mathit{logit}(\pi(x)) = \log \frac{\pi(x)}{1-\pi(x)}$ has linear pattern with respect to x from our observed data. So given x, we model $\mathit{logit}(\pi(x))$ as a linear function of x, that is

$$\mathit{logit}(\pi(x)) = \beta_0 + \beta_1 x.$$

- Therefore, the probability model for simple logistic regression can be represented as follows,

$$\text{Given } x,\ y \sim \mathit{Binomial}(1, \pi(x)),$$

$$\mathit{logit}(\pi(x)) = \log \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x.$$

- The sign of β_1 determines whether $\pi(x)$ is increasing or decreasing as x increases. If $\beta_1 = 0$, y is independent of x.

Logistic Regression

2. **Estimate model parameters (β_0, β_1)**

- The probability of observing y_i given x_i known:

$$\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}.$$

  - If $y_i = 1$, then the chance is $\pi(x_i)$;
  - If $y_i = 0$, then the chance is $1 - \pi(x_i)$.

- The likelihood function of the observed $(y_1, y_2, ..., y_n)$:

$$\prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}, \textit{where } \pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.$$

- Indeed, the likelihood function is a function of β_0, β_1 and we can denote it by $L(\beta_0, \beta_1)$.

Logistic Regression

- **Maximum likelihood estimation**: The likelihood function $L(\beta_0, \beta_1)$ defined above actually measures the probability of observing $(y_1, ..., y_n)$. So the best estimators for β_0, β_1 would be the ones that maximize the likelihood function, that is, finding $\hat{\beta}_0$ and $\hat{\beta}_1$ such that

$$L(\hat{\beta}_0, \hat{\beta}_1) = max_{\beta_0, \beta_1} L(\beta_0, \beta_1).$$

Logistic Regression

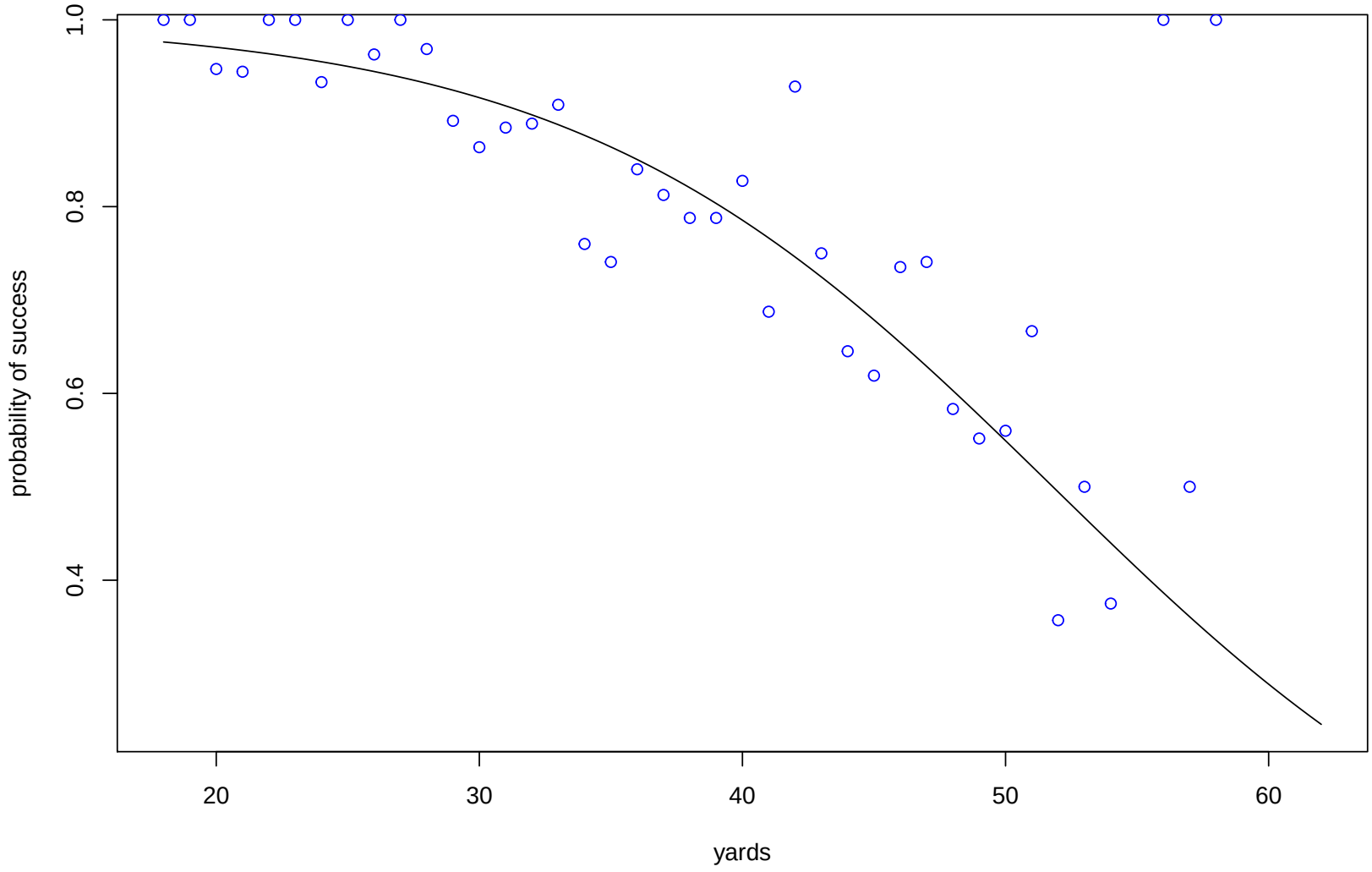
Use the following R code we can get $\hat{\beta}_0, \hat{\beta}_1$.

```
data <- read.table("fieldgoal.dat",head=F)
y <- data[,2]
x <- data[,1]
myglm <- glm(y ~ x, family=binomial("logit"))
#the estimated beta0, beta1
beta0 <- round(coef(myglm)[1],3)
beta1 <- round(coef(myglm)[2],3)
c(beta0,beta1)
```

```
## (Intercept)           x
##       5.698      -0.110
```

Logistic Regression

The fitted probability of success $\hat{\pi}(x) = \frac{e^{5.698-0.11x}}{1+e^{5.698-0.11x}}$.



Logistic Regression

3. Inference

- Q1: In practice, $\frac{\pi(x)}{1-\pi(x)}$ is called **odds ratio**. The logistic regression basically assume the log of odds ratio has linear pattern with respect to explanatory variable x. In **Example 9.1**, is the log of odds ratio a linear function of the attempt distance?
  - Consider the test

$$H_0 : \beta_1 = 0 \text{ vs. } H_a : \beta_1 \neq 0.$$

  - When the sample size is large, we could use Wald test, that is

$$z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim Normal(0, 1) \text{ if } H_0 \text{ is true,}$$

    where $se(\hat{\beta}_1)$ means the standard error of β_1.

Logistic Regression

- We can find $se(\hat{\beta}_1)$, z value and even the p-value using R software,

```
round(summary(myglm)$coefficients,3)
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    5.698      0.451  12.631        0
## x             -0.110      0.011 -10.385        0
```

From the R output, we can see the observed z value is -10.385 and hence the p value is almost zero. So we reject the null hypothesis, which means there is strong evidence that the log of odds ratio is a linear function of the attempt distance.

Logistic Regression

- Q2: Given $x_0 = 40$, construct the 95% confidence interval for the probablity of success $\pi(x_0)$.
  - When sample size is large, the 95% confidence interval for log of odds ratio $\mathit{logit}(\pi(x_0))$:

$$(c_1, c_2) := (\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm 1.96 \times se(\hat{\beta}_0 + \hat{\beta}_1 x_0),$$

    where $se(\hat{\beta}_0 + \hat{\beta}_1 x_0)$ is the standard error of $\hat{\beta}_0 + \hat{\beta}_1 x_0$.
  - So the 95% confidence interval for the probability of success $\pi(x_0)$ is

$$\left(\frac{e^{c_1}}{1 + e^{c_1}}, \frac{e^{c_2}}{1 + e^{c_2}} \right).$$

Logistic Regression

- Using R, we can find out $\hat{\pi}(x_0)$ and $se(\hat{\pi}(x_0))$, and then construct the 95% confidence interval for the probability of success $\pi(x_0)$ when $x_0 = 40$.

```
pred.glm <-  predict.glm(myglm, data.frame(x = 40),
                                se.fit=TRUE)
#95% CI for odds ration
OR <- c(pred.glm$fit-1.96*pred.glm$se.fit,
         pred.glm$fit+1.96*pred.glm$se.fit)
#95% CI for pi(x)
CI.pi <- round(exp(OR)/(1+exp(OR)),4)
names(CI.pi) <- c("2.5th", "97.5th")
CI.pi
```

```
##  2.5th 97.5th
## 0.7551 0.8141
```