Lecture 5: Logistic regression & NFL kickers

Skidmore College

Goals

- Kicker statistics, NFL
- Extensions: Expected points
- Tools: Logistic regression

Review: multivariate linear regression

Model:

$$y_i = \beta_0 + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \ldots + \beta_{p-1} * x_{i,p-1} + \epsilon_i$$

Assumptions:

- $\epsilon_i \sim N(0, \sigma^2)$
- ϵ_i,$\epsilon_{i'}$ independent for all i, i'
- Linear relationship between y and x

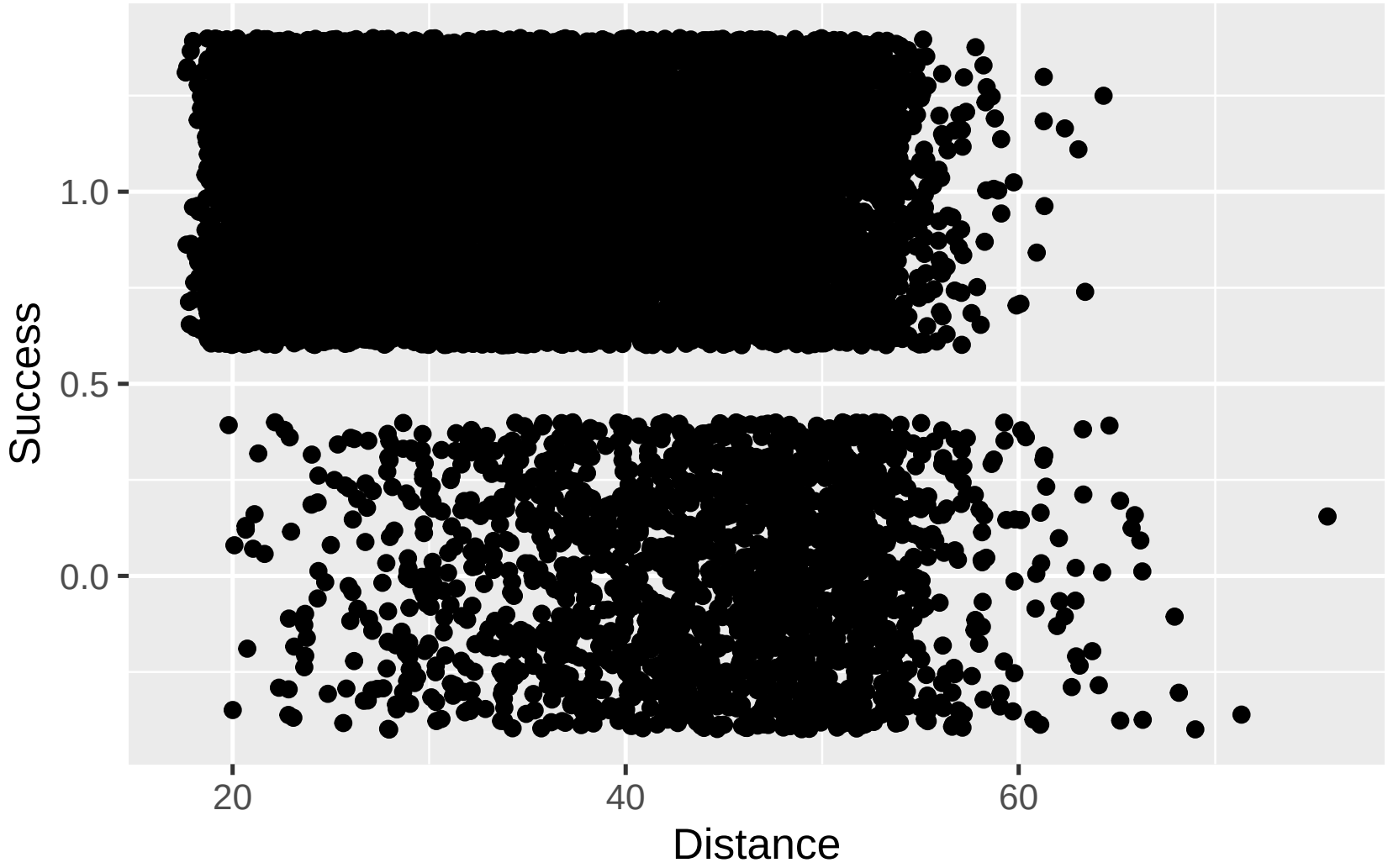
Example: NFL kickers

```
library(RCurl); library(tidyverse)
url <- getURL("https://raw.githubusercontent.com/statsbylopez/StatsSports/maste
nfl_kick <- read.csv(text = url)
head(nfl_kick)
```

```
##   Team Year GameMinute Kicker Distance ScoreDiff Grass Temp Success
## 1  PHI 2005          3  Akers       49         0 FALSE   72       0
## 2  PHI 2005         29  Akers       49        -7 FALSE   72       0
## 3  PHI 2005         51  Akers       44        -7 FALSE   72       1
## 4  PHI 2005         14  Akers       43        14  TRUE   82       0
## 5  PHI 2005         60  Akers       23         0  TRUE   75       1
## 6  PHI 2005         39  Akers       34        -3  TRUE   68       1
```

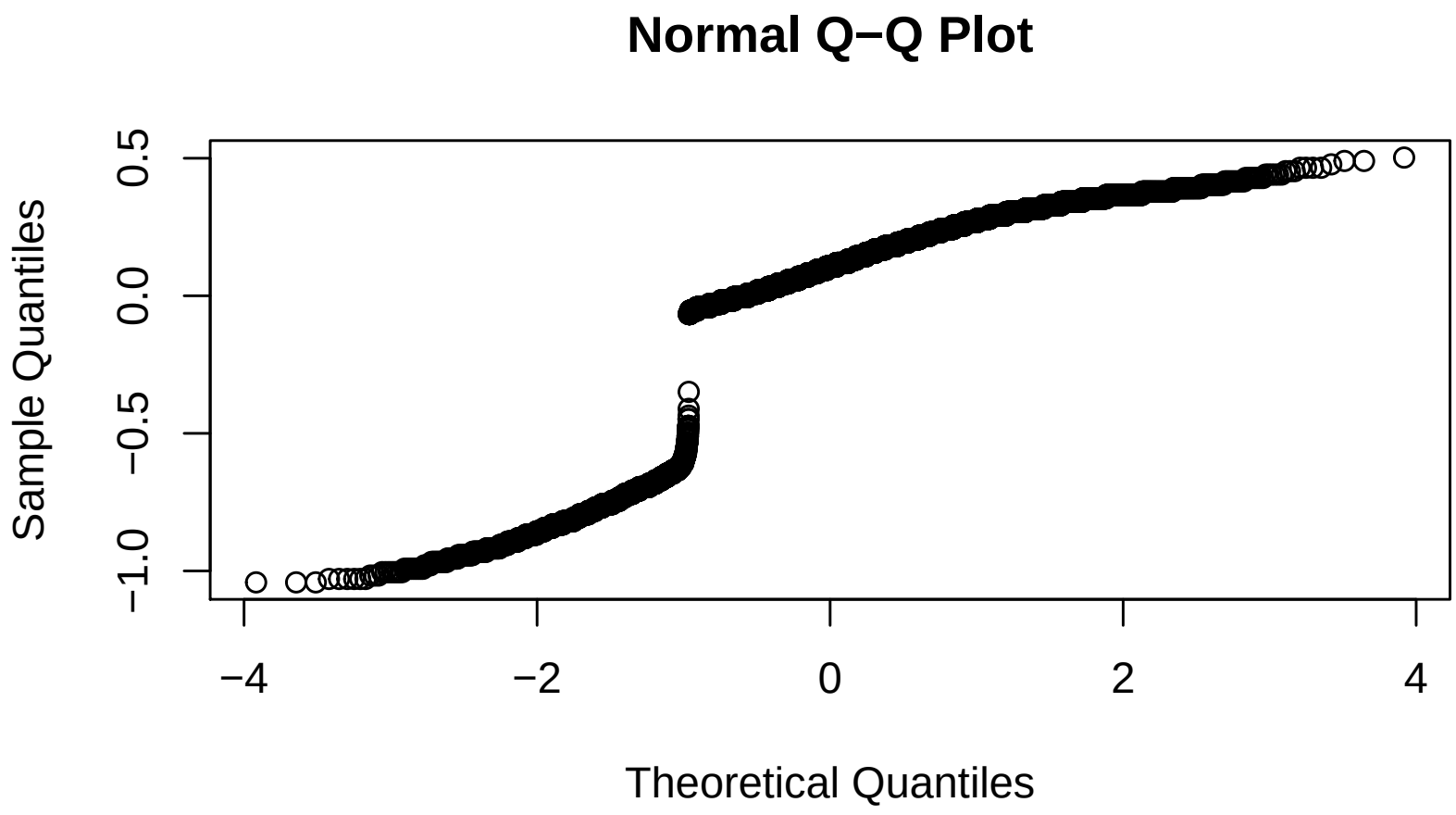
Example: NFL kickers

```
fit_0 <- lm(Success ~ Distance, data = nfl_kick)
ggplot(data = nfl_kick, aes(Distance, Success)) +
  geom_jitter()
```



Example: NFL kickers

```
fit_0 <- lm(Success ~ Distance, data = nfl_kick)
qqnorm(fit_0$resid)
```



What are the problems?

Logistic regression model

Model: $log(\frac{P(y=1)}{1-P(y=1)}) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \ldots + \beta_{p-1} * x_{p-1}$

Comments:

- Dependent variable: log-odds
  - What are odds?
- Model checks more complex
- Uses z test statistics for parameters

Logistic regression model

Model: $log(\frac{P(y=1)}{1-P(y=1)}) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2$

Extract probabilities:

- $P(y = 1)$:

Estimated logistic regression model

Estimated model:
$log(\frac{P(y=1)}{1-P(y=1)}) = \hat{\beta}_0 + \hat{\beta}_1 * x_1 + \hat{\beta}_2 * x_2 + \ldots + \hat{\beta_{p-1}} * x_{p-1}$

Slope interpretation:

- $\hat{\beta}_1$:
- $e^{\hat{\beta}_1}$:

Ex: Field goal kicking by distance

Model: $log(\frac{P(Success=1)}{1-P(Success=1)}) = \beta_0 + \beta_1 * Distance$

```
library(broom)
fit_1 <- glm(Success ~ Distance, data = nfl_kick, family = "binomial")
tidy(fit_1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    5.72    0.137        41.7 0.
## 2 Distance      -0.103   0.00314     -32.7 5.63e-235
```

Slope interpretation: $e^{\hat{\beta}_1}$

Ex: Field goal kicking by distance

```
tidy(fit_1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    5.72    0.137        41.7 0.
## 2 Distance      -0.103   0.00314     -32.7 5.63e-235
```

Estimate the probability of a successful 50-yard field goal:

Ex: Field goal kicking by distance

```
tidy(fit_1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    5.72    0.137        41.7 0.
## 2 Distance      -0.103   0.00314     -32.7 5.63e-235
```

Estimate the probability of a successful 51-yard field goal:

Ex: Field goal kicking by distance

Use your answers on the previous slides to estimate the odds of a 51-yard field goal relative to the odds of a 50-yard field goal. Where else do you see this number?

Model checking

- Model checking for logistic regression relies on assessment of fit
  - Are the predicted probabilities accurate?
  - Ex: 48 to 52 yard field goals

```
long_FG <- filter(nfl_kick, Distance >= 48, Distance <= 52)
long_FG %>%
  summarise(ave_success = mean(Success))
```

```
##   ave_success
## 1   0.6510989
```

Expected points

Probability: $E(X) = \sum_{i=1} x_i * P(X = x_i)$

1. What is expected points from a 50 yard field goal?
2. What are expected points for 20, 30, 40, 50, 60 yard field goals?

```
Distance <- data.frame(Distance = c(20, 30, 40, 50, 60))
pred <- predict(fit_1, Distance, type = "response")
round(pred, 2)*3
```

```
##    1    2    3    4    5
## 2.94 2.79 2.49 1.92 1.17
```

Why it matters?

Offensive decision making:

1. Maximize number of points scored on a drive
   - What does this entail?
   - Should we kick the field goal or punt?
   - Should we kick the field goal or go for it?
   - When to go for it on fourth down?
2. Maximize chances of winning the game
   - When does this differ from (1)?