Lecture 19: Logistic Regression

Module 6: part 1

Spring 2025

Logistics

- Starting Module 6 today on generalized linear models
- Module 5 assessment due Wed Apr 22

Recap

BTRY 6020 so far . . .

- So far we've learned a lot about linear models
- **Module 1**: simple linear regression with 1 covariate:

$$E(Y \mid X = x) = b_0 + bx \qquad \text{or} \qquad Y_i = b_0 + bX_i + \varepsilon_i$$

- We can compute $\hat{b}_0$ and $\hat{b}_1$ to estimate b_0 and b_1 by minimizing

$$RSS = \sum_i (y_i - \hat{y}_i)^2 \qquad \text{where} \qquad \hat{y}_i = \hat{b}_0 + \hat{b}_1 x_i$$

BTRY 6020 so far . . .

- **Module 2**: extended the framework to include multiple covariates

$$E(Y \mid \mathbf{X} = \mathbf{x}) = b_0 + \sum_{k=1}^{p} b_p x_p \qquad \text{or} \qquad Y_i = b_0 + \sum_{k=1}^{p} b_k X_{i,k} + \varepsilon_i$$

- Now, b_k represents the associated difference in the expected value of Y when comparing two observations who's X_k values differ by 1 unit, but all other covariates are the same
- Can flexibly model $E(Y_i \mid \mathbf{X_i} = \mathbf{x})$, the conditional mean of Y_i given $\mathbf{X_i}$
  - Can control for other variables
  - Can include categorical variables as dummy terms
  - Can include polynomial terms
  - Can use transformations of the covariates and the dependent variable

Housing example:

- **Module 2**: extended the framework to include multiple covariates

$$E(Y \mid \mathbf{X} = \mathbf{x}) = b_0 + \sum_{k=1}^{p} b_p x_p \qquad \text{or} \qquad Y_i = b_0 + \sum_{k=1}^{p} b_k X_{i,k} + \varepsilon_i$$

- Now, b_k represents the associated difference in the expected value of Y when comparing two observations who's X_k values differ by 1 unit, but all other covariates are the same
- Can flexibly model $E(Y_i \mid \mathbf{X_i} = \mathbf{x})$, the conditional mean of Y_i given $\mathbf{X_i}$
  - Can control for other variables
  - Can include categorical variables as dummy terms
  - Can include polynomial terms
  - Can use transformations of the covariates and the dependent variable

Testing in linear models

Module 3: Hypothesis Testing

- Use a T-test to test a single coefficient
- Use a F-test to test multiple coefficients simultaneously

Module 4: How can we still do testing when the assumptions are violated

- When the data generating procedure is heteroscedastic
- Bootstrap can be a powerful tool for estimating the standard errors that doesn't require as many assumptions
- When different observations may not be independent of each other, then use fixed effects or random effects

Module 5: How to choose which covariates to include when you have many to consider

One major restriction

Up until now, we've always assumed that our dependent variable Y is continuous (or at least close enough that we can model it that way)

One major restriction

Up until now, we've always assumed that our dependent variable Y is continuous (or at least close enough that we can model it that way)

How can we model data that is discrete or count data?

Modeling discrete data

Example: Modeling NFL field goals

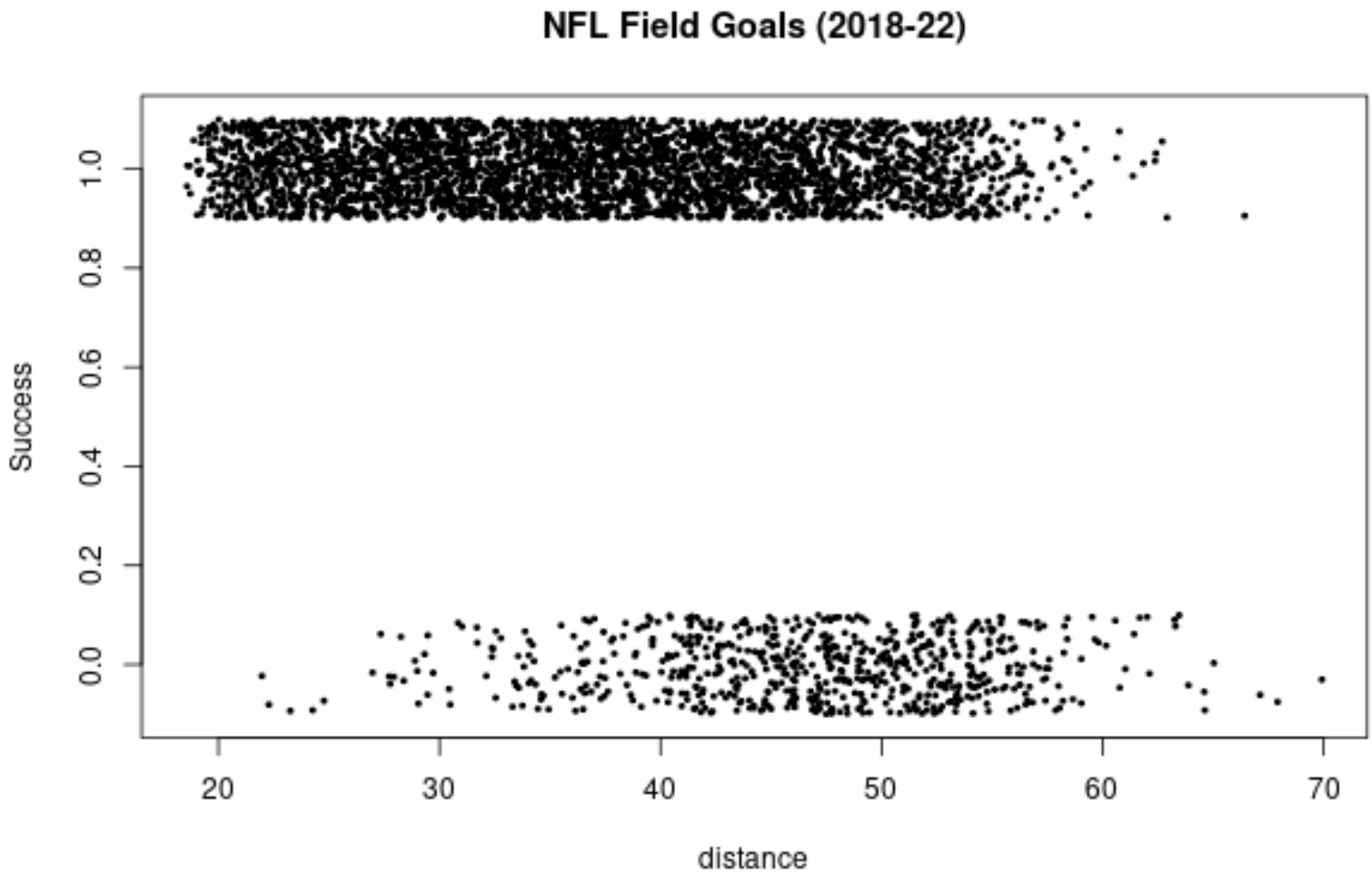
In American football, if you can kick the football through the field goal you get three points

Example: Modeling NFL field goals

In American football, if you can kick the football through the field goal you get three points

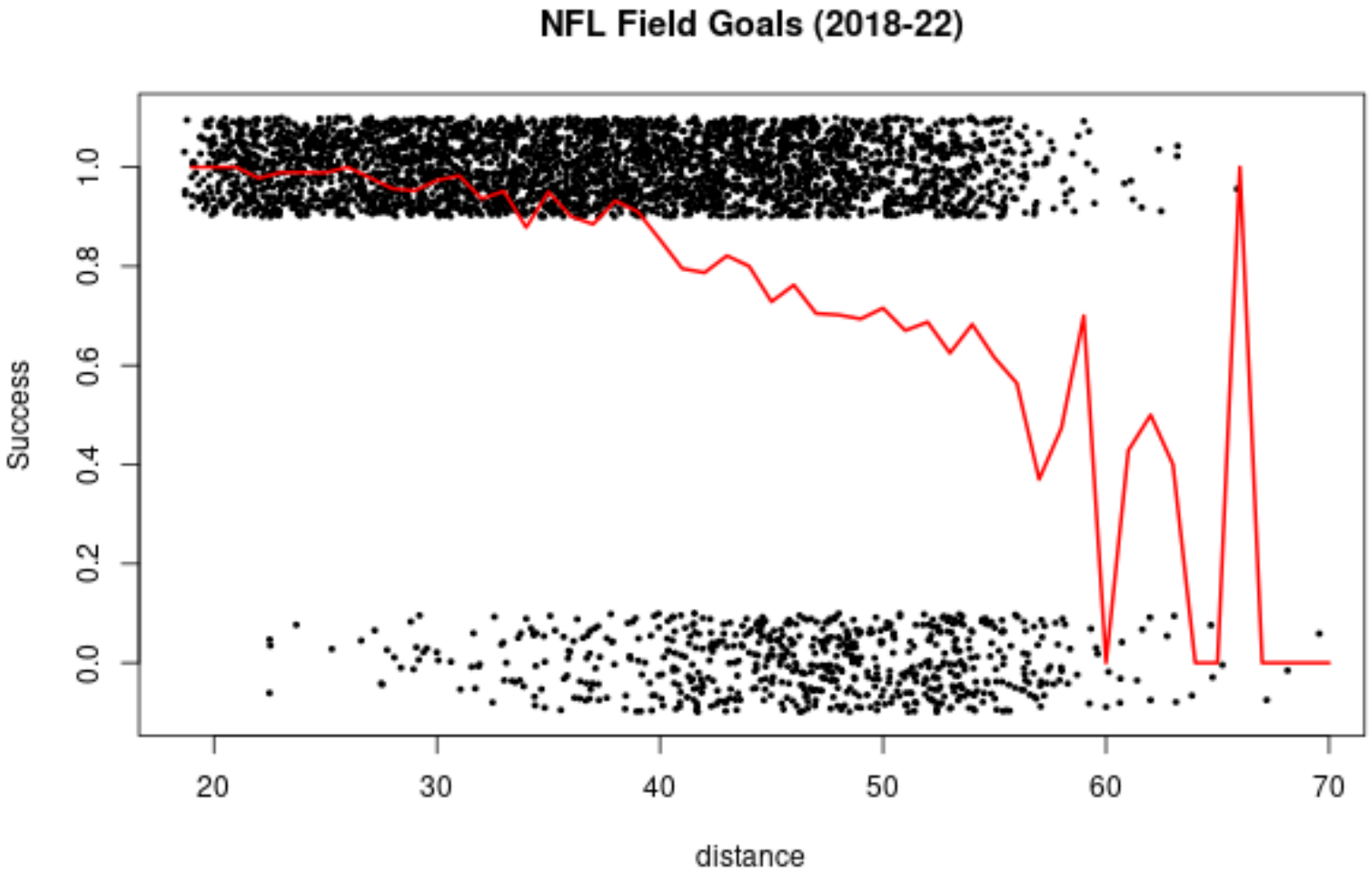
Example: Modeling NFL field goals

In American football, if you can kick the football through the field goal you get three points



Example: Modeling NFL field goals

In American football, if you can kick the football through the field goal you get three points



Example: Modeling NFL field goals

If we regress the outcome of the kick where Miss $= 0$ and Make $= 1$ onto

- Distance (yards)
- Wind Speed (mph)
- Raining $= 1$, Dry $= 0$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.3873	0.0333	41.62	0.0000
distance	-0.0136	0.0008	-17.65	0.0000
Wind Speed	-0.0042	0.0016	-2.57	0.0103
Rain	-0.0509	0.0347	-1.47	0.1419

Example: Modeling NFL field goals

If a kick is from 35 yards, the wind speed is 10 mph, and it is raining, then we would predict that

$$Y_i = 1.388 - .014 \times (35) - .004 \times (10) - .051(1) = .877$$

Example: Modeling NFL field goals

If a kick is from 35 yards, the wind speed is 10 mph, and it is raining, then we would predict that

$$Y_i = 1.388 - .014 \times (35) - .004 \times (10) - .051(1) = .877$$

What model are we actually assuming?

$$Y_i = b_0 + \sum_{k=1}^{p} b_k X_k + \varepsilon_i$$

The range of possible ε_i depends on $b_0 + \sum_{k=1}^{p} b_k X_k$

Bernoulli Distribution

Bernoulli Distribution is used to model binary variables

- Suppose a random variable Y has outcomes $\{0, 1\}$
- We only need to specify the parameter $\theta = P(Y = 1)$ because $P(Y = 0) = 1 - \theta$
- The parameter $0 \leq \theta \leq 1$
- For Y, we have $E(Y) = \theta$ and $\text{var}(Y) = \theta(1 - \theta)$

Example: Modeling NFL field goals

We can estimate the probability of success as a linear function of covariates

$$P(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \theta(\mathbf{x}) = E(Y \mid \mathbf{X} = \mathbf{x}) = b_0 + \sum_{k=1}^{p} b_k x_k$$

If a kick is from 35 yards, the wind speed is 10 mph, and it is raining, then the probability of success is

$$\theta(\mathbf{x}) = 1.388 - .014 \times (35) - .004 \times (10) - .051(1) = .877$$

Example: Modeling NFL field goals

We can estimate the probability of success as a linear function of covariates

$$P(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \theta(\mathbf{x}) = E(Y \mid \mathbf{X} = \mathbf{x}) = b_0 + \sum_{k=1}^{p} b_k x_k$$

If a kick is from 35 yards, the wind speed is 10 mph, and it is raining, then the probability of success is

$$\theta(\mathbf{x}) = 1.388 - .014 \times (35) - .004 \times (10) - .051(1) = .877$$

If a kick is from 10 yards, the wind speed is 5 mph, and it is not raining, then the probability of success is

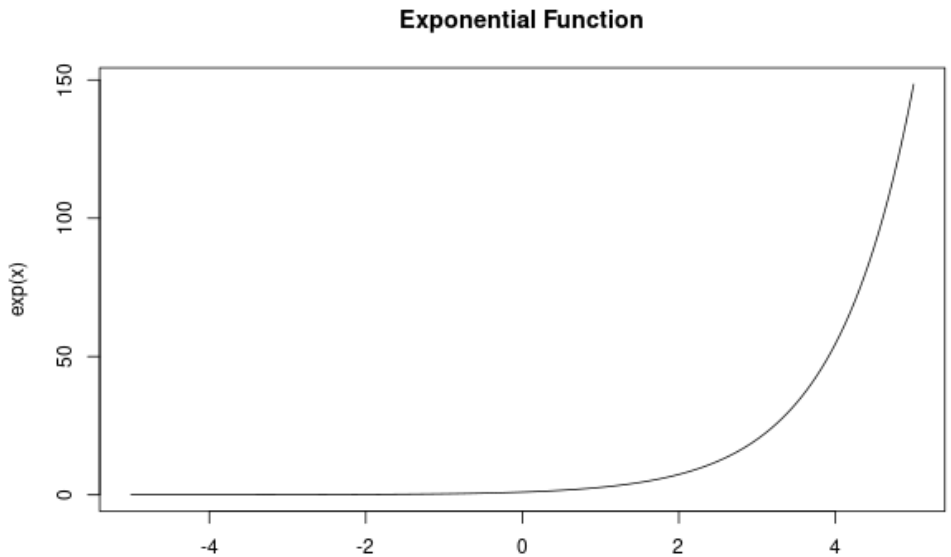
$$\theta(\mathbf{x}) = 1.388 - .014 \times (10) - .004 \times (5) - .051(0) = 1.08$$

Modeling the probability of success

- We want a function who's input $(b_0 + \sum_{k=1}^{p} b_k x_k)$ can be any value $(-\infty, \infty)$, but the output is $(0, 1)$
- We use the logistic function

$$s(z) = \frac{\exp(z)}{1 + \exp(z)}$$

- When z is very small (i.e., very negative), the numerator is very close to 0, so $s(z) \approx 0$
- When z is very large, the numerator and the denominator are both very large so $s(z) \approx 1$

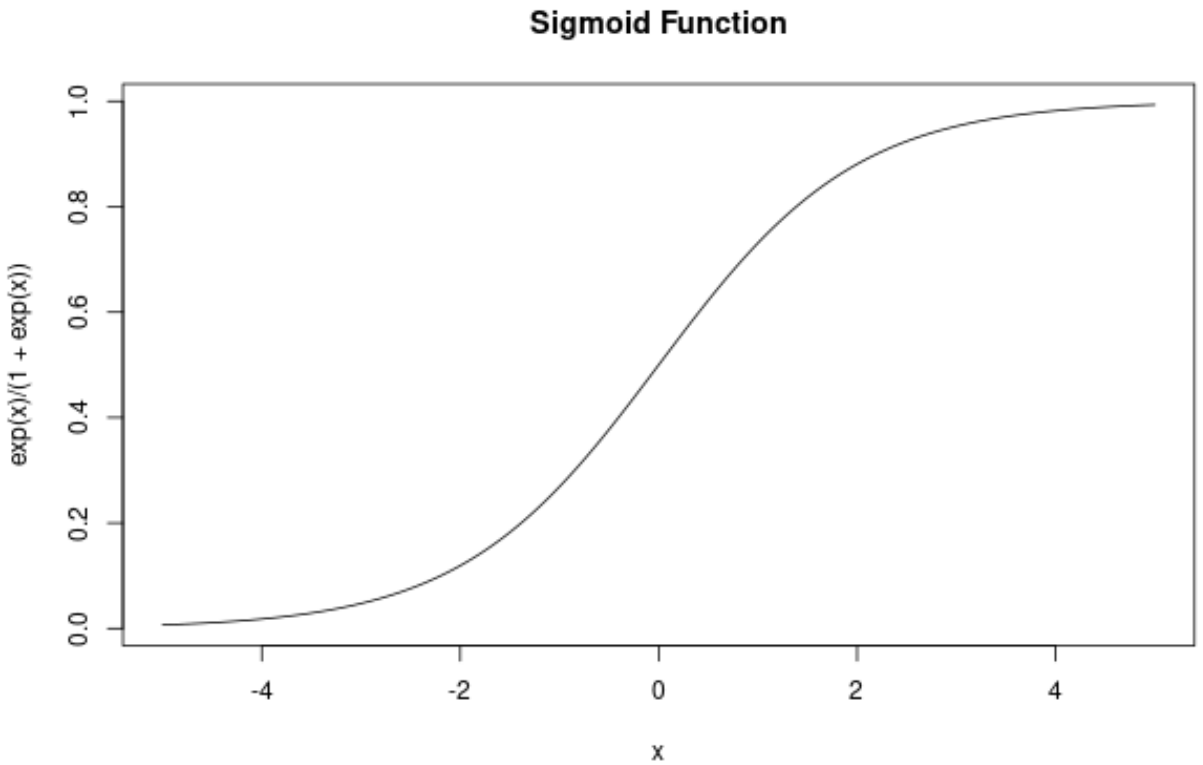


Modeling the probability of success

So we can fit a model such that

$$P(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \theta(\mathbf{x}) = E(Y \mid \mathbf{X} = \mathbf{x}) = \frac{\exp(b_0 + \sum_{k=1}^{p} b_k x_k)}{1 + \exp(b_0 + \sum_{k=1}^{p} b_k x_k)}$$

- Stays between $(0, 1)$
- Diminishing returns



Logistic Regression

This is equivalent to

$$\log\left(\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})}\right) = b_0 + \sum_{k=1}^{p} b_k x_k$$

- The function $\log(\theta/(1-\theta))$ is called the logit function
- This model is called **Logistic regression**
- $\theta/(1-\theta)$ are called the odds
- Log odds are a linear function of the covariates

Logistic Regression

This is equivalent to

$$\log\left(\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})}\right) = b_0 + \sum_{k=1}^{p} b_k x_k$$

- The function $\log(\theta/(1-\theta))$ is called the logit function
- This model is called **Logistic regression**
- $\theta/(1-\theta)$ are called the odds; can range from $(0, \infty)$
- Log odds are a linear function of the covariates; can range from $(-\infty, \infty)$

Logistic Regression

$$\log\left(\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})}\right) = b_0 + \sum_{k=1}^{p} b_k x_k$$

- Suppose we set all $x_k = 0$

$$\log\left(\frac{\theta(\mathbf{0})}{1-\theta(\mathbf{0})}\right) = b_0$$

- The intercept is the value of the log-odds when all covariates are 0
- May not be meaningful if covariates can never be 0

Logistic Regression

$$\log\left(\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})}\right) = b_0 + \sum_{k=1}^{p} b_k x_k$$

- Suppose $\mathbf{x_1}$ and $\mathbf{x_2}$ are individuals whose covariates values which are the all the same, except that $x_{2,p} = x_{1,p} + 1$

$$\begin{aligned}
&\log\left(\frac{\theta(\mathbf{x_2})}{1-\theta(\mathbf{x_2})}\right) - \log\left(\frac{\theta(\mathbf{x_1})}{1-\theta(\mathbf{x_1})}\right) \\
&\quad = b_0 + \sum_{k=1}^{p-1} b_k x_{2,k} b_0 + b_p x_{2,p} - b_0 - \sum_{k=1}^{p-1} b_k x_{1,k} - b_p x_{1,p} \\
&\quad = b_p(x_{2,p} - x_{1,p}) = b_p
\end{aligned}$$

Logistic Regression

By properties of the log

$$\log\left(\frac{\theta(\mathbf{x_2})}{1-\theta(\mathbf{x_2})}\right) - \log\left(\frac{\theta(\mathbf{x_1})}{1-\theta(\mathbf{x_1})}\right) = \log\left(\frac{\theta(\mathbf{x_2})/(1-\theta(\mathbf{x_2}))}{\theta(\mathbf{x_1})/(1-\theta(\mathbf{x_1}))}\right)$$

so putting everything together, we have

$$\frac{\theta(\mathbf{x_2})/(1-\theta(\mathbf{x_2}))}{\theta(\mathbf{x_1})/(1-\theta(\mathbf{x_1}))} = \exp(b_p) \tag{1}$$

Logistic Regression

By properties of the log

$$\log\left(\frac{\theta(\mathbf{x_2})}{1-\theta(\mathbf{x_2})}\right) - \log\left(\frac{\theta(\mathbf{x_1})}{1-\theta(\mathbf{x_1})}\right) = \log\left(\frac{\theta(\mathbf{x_2})/(1-\theta(\mathbf{x_2}))}{\theta(\mathbf{x_1})/(1-\theta(\mathbf{x_1}))}\right)$$

so putting everything together, we have

$$\frac{\theta(\mathbf{x_2})/(1-\theta(\mathbf{x_2}))}{\theta(\mathbf{x_1})/(1-\theta(\mathbf{x_1}))} = \exp(b_p) \tag{1}$$

- Odds ratio: $\frac{\theta(\mathbf{x_2})/(1-\theta(\mathbf{x_2}))}{\theta(\mathbf{x_1})/(1-\theta(\mathbf{x_1}))}$
- **Interpretation**: If observation 1 and observation 2 have all the same covariates, but $x_{2,p} = x_{1,p} + 1$, then the odds for Y_2 are $\exp(b_p)$ times larger (i.e., multiplicative) than the odds for Y_1

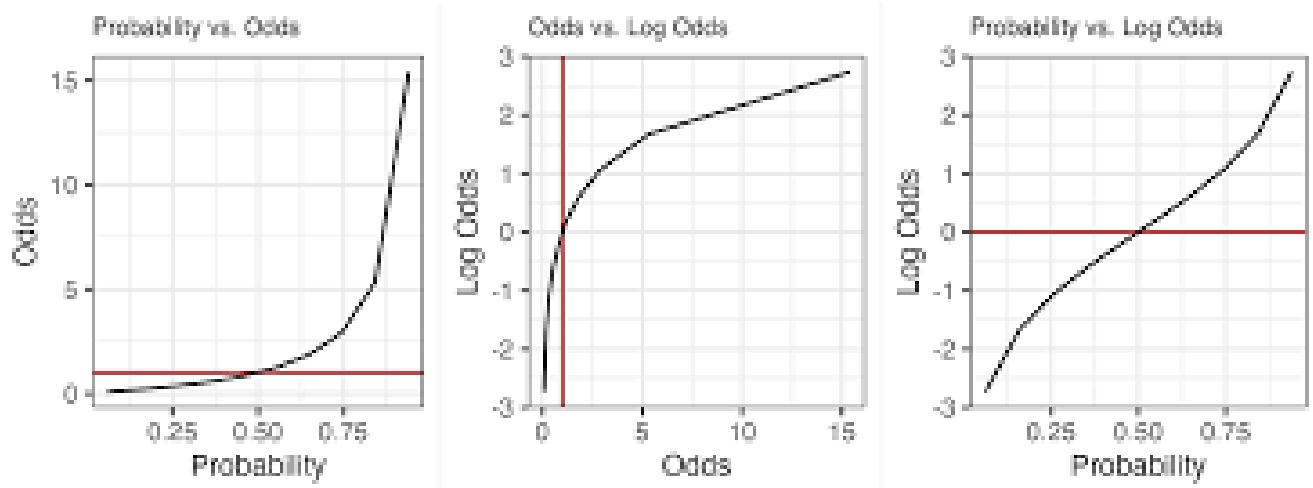
Odds and Odds ratios

The odds and odds ratios are a bit difficult to interpret concretely

- When θ is very small, the odds are close to the probability of success

$$\frac{\theta}{1-\theta} \approx \frac{\theta}{1} = \theta$$

- Can always map the odds back to the probability $\theta = \frac{\text{odds}}{1+\text{odds}}$
- Can always map the log-odds back to the probability $\theta = \frac{\exp(\text{log odds})}{1+\exp(\text{log odds})}$
- Odds and probability always move in the same direction (i.e., increasing/decreasing one always increases/decreases the other)



Odds and Odds ratios

The odds and odds ratios are a bit difficult to interpret concretely

- When the odds ratio (often abbreviated as OR) of Y_2 vs Y_1 is > 1, then $P(Y_2 = 1) > P(Y_1 = 1)$
- When $OR = 1$ then $P(Y_2 = 1) = P(Y_1 = 1)$
- When $OR < 1$ then $P(Y_2 = 1) < P(Y_1 = 1)$

Odds and Odds ratios

The odds and odds ratios are a bit difficult to interpret concretely

- When the odds ratio (often abbreviated as OR) of Y_2 vs Y_1 is > 1, then $P(Y_2 = 1) > P(Y_1 = 1)$
- When $OR = 1$ then $P(Y_2 = 1) = P(Y_1 = 1)$
- When $OR < 1$ then $P(Y_2 = 1) < P(Y_1 = 1)$

When Observation 2 and Observation 1 have all the same covariates except, $x_{2,p} = x_{1,p} + 1$, then the odds ratio of Y_2 vs Y_1 is $\exp(b_p)$

- When $b_p > 0$ then $OR > 1$
- When $b_p = 0$ then $OR = 1$
- When $b_p < 0$ then $OR < 1$

so the coefficients sign (positive or negative) indicates whether larger values of X_p are associated with a higher probability of success

NFL Field Goals

We use logistic regression to model the log odds of a successful kick as a linear function of

- Distance (yards)
- Wind Speed (mph)
- Raining $= 1$, Dry $= 0$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	6.8185	0.3823	17.84	0.0000
Distance	-0.1174	0.0079	-14.91	0.0000
Wind Speed	-0.0355	0.0128	-2.77	0.0056
Rain	-0.4385	0.2613	-1.68	0.0933

NFL Field Goals

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	6.8185	0.3823	17.84	0.0000
Distance	-0.1174	0.0079	-14.91	0.0000
Wind Speed	-0.0355	0.0128	-2.77	0.0056
Rain	-0.4385	0.2613	-1.68	0.0933

- Considering two attempts with the same rain and wind conditions, the odds of a successful attempt of a kick are $\exp(-.1174) = .889$ of the odds of a kick which is 1 yard longer
- Considering two attempts with the same distance and wind speed, when it is raining, the odds of a successful attempt are $\exp(-.439) = .644$ of the odds when it is not raining

NFL Field Goals

If a kick is from **35 yards**, the wind speed is 10 mph, and it is not raining, then we estimate that

$$\log\left(\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})}\right) = 6.819 - .117 \times (35) - .036 \times (10) - .439(0) = 2.364$$

$$\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})} = \exp(2.364) = 10.6334$$

$$P(Success) = \theta(\mathbf{x}) = \frac{\exp(2.364)}{1+\exp(2.364)} = .914$$

If a kick is from **36 yards**, the wind speed is 10 mph, and it is not raining, then we estimate that

$$\log\left(\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})}\right) = 6.819 - .117 \times (36) - .036 \times (10) - .439(0) = 2.247$$

$$\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})} = \exp(2.247) = 9.459 = \exp(2.364) \times .889$$

$$P(Success) = \theta(\mathbf{x}) = \frac{\exp(2.247)}{1+\exp(2.247)} = 0.904$$

NFL Field Goals

If a kick is from 35 yards, the wind speed is 10 mph, and it **is not** raining, then we estimate that

$$\log\left(\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})}\right) = 6.819 - .117 \times (35) - .036 \times (10) - .439(0) = 2.364$$

$$\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})} = \exp(2.364) = 10.6334$$

$$P(Success) = \theta(\mathbf{x}) = \frac{\exp(2.364)}{1+\exp(2.364)} = .914$$

If a kick is from 35 yards, the wind speed is 10 mph, and it **is** raining, then we estimate that

$$\log\left(\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})}\right) = 6.819 - .117 \times (35) - .036 \times (10) - .439(1) = 1.935$$

$$\frac{\theta(\mathbf{x})}{1-\theta(\mathbf{x})} = \exp(1.935) = 6.855 = \exp(2.364) \times .644$$

$$P(Success) = \theta(\mathbf{x}) = \frac{\exp(1.935)}{1+\exp(1.935)} = .873$$

Wrap-up

- Modeling discrete data requires different approach
- Model parameter (or transformation of parameter) used linear model
- For binary data, we model log-odds
- Interpret model using odds ratio