STAT 213
Logistic Regression

Colin Reimer Dawson

Oberlin College

July 26, 2021

Quantitative Vs. Categorical Predictor and Response

		Response	
		Quantitative	**Categorical**
Predictor	**Quantitative**	Linear Reg.	Logistic Reg.
	Categorical	ANOVA	

Binary Logistic Regression

Response variable (Y) is categorical with two categories (i.e., binary).

- Code Y as an indicator variable: 0 or 1
- Assume (for now) a single quantitative predictor, X

Three Equivalent Forms of Logistic Regression

Probability Form

$$\pi := \text{Probability}(Y = 1) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Odds Form

$$\text{Odds}(Y = 1) = \frac{\pi}{1 - \pi} = e^{\beta_0 + \beta_1 X}$$

Logit Form

$$\eta := \text{LogOdds}(\text{Y} = 1) = \log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 X$$

Example: Golf Putts

Distance (ft)	3	4	5	6	7
# Made	84	88	61	61	44
# Missed	17	31	47	64	90
Total	101	119	108	125	134

1. Estimate the *probability* of success at each length
2. Estimate the *odds* of success at each length
3. Estimate the *log odds* of success at each length

Golf Putts: Solutions

Interpretations

- **Probability**: The long run ratio of "positive" instances to "total" instances
  - Scale: 0 to 1
  - Both outcomes equally likely: Probability $= 1/2$
- **Odds**: Long run ratio of "positive" to "negative" instances
  - Scale: 0 to ∞
  - Both outcomes equally likely: Odds $= 1$
- **Log Odds**: Log of the odds
  - Scale: $-\infty$ to ∞
  - Both outcomes equally likely: Log Odds $= 0$
  - Nice features of this scale:
    - Its range is unrestricted
    - Swapping the probability of success and failure just switches the sign of the log odds

Odds Ratios

Logit and Odds

$$\begin{aligned} \textbf{Logit/Log Odds:} \quad \log\left(\frac{\pi}{1-\pi}\right) &= \beta_0 + \beta_1 X \\ \textbf{Odds:} \quad \frac{\pi}{1-\pi} &= e^{\beta_0 + \beta_1 X} \end{aligned}$$

- In the model, for each 1 unit increase in X, the **logit/log odds** increases by β_1.
- Equivalently: For each 1 unit increase in X, the **odds** are **multiplied** by e^{β_1}
- e^{β_1} is the **odds ratio** comparing $P(Y=1)$ at $X+1$ compared to $P(Y=1)$ at X:

$$\frac{e^{\beta_0+\beta_1(X+1)}}{e^{\beta_0+\beta_1 X}} = \frac{e^{\beta_0}e^{\beta_1 X}e^{\beta_1}}{e^{\beta_0}e^{\beta_1 X}} = e^{\beta_1} \text{ regardless of } X$$

Odds Ratios

The **odds ratio** associated with a binary response Y at two different predictor values $X = x_1$ vs. $X = x_2$ is the ratio of the odds at $X = x_2$ to the odds at $X = x_1$.

Denote by $\pi(x)$ the probability of $Y = 1$ when $X = x$. Then

$$\text{Odds Ratio}(x_2 \text{ vs. } x_1) = \frac{\pi(x_2)/(1 - \pi(x_2))}{\pi(x_1)/(1 - \pi(x_1))}$$

We can estimate this from a sample using:

$$\widehat{\text{Odds Ratio}}(x_2 \text{ vs. } x_1) = \frac{\hat{\pi}(x_2)/(1 - \hat{\pi}(x_2))}{\hat{\pi}(x_1)/(1 - \hat{\pi}(x_1))}$$

Example: Golf Putts

Distance (ft)	3	4	5	6	7
# Made	84	88	61	61	44
# Missed	17	31	47	64	90
Total	101	119	108	125	134
$\hat{\pi}$	0.832	0.739	0.565	0.488	0.328
Odds	4.94	2.84	1.30	0.95	0.49
Log Odds	1.60	1.04	0.26	-0.05	-0.71

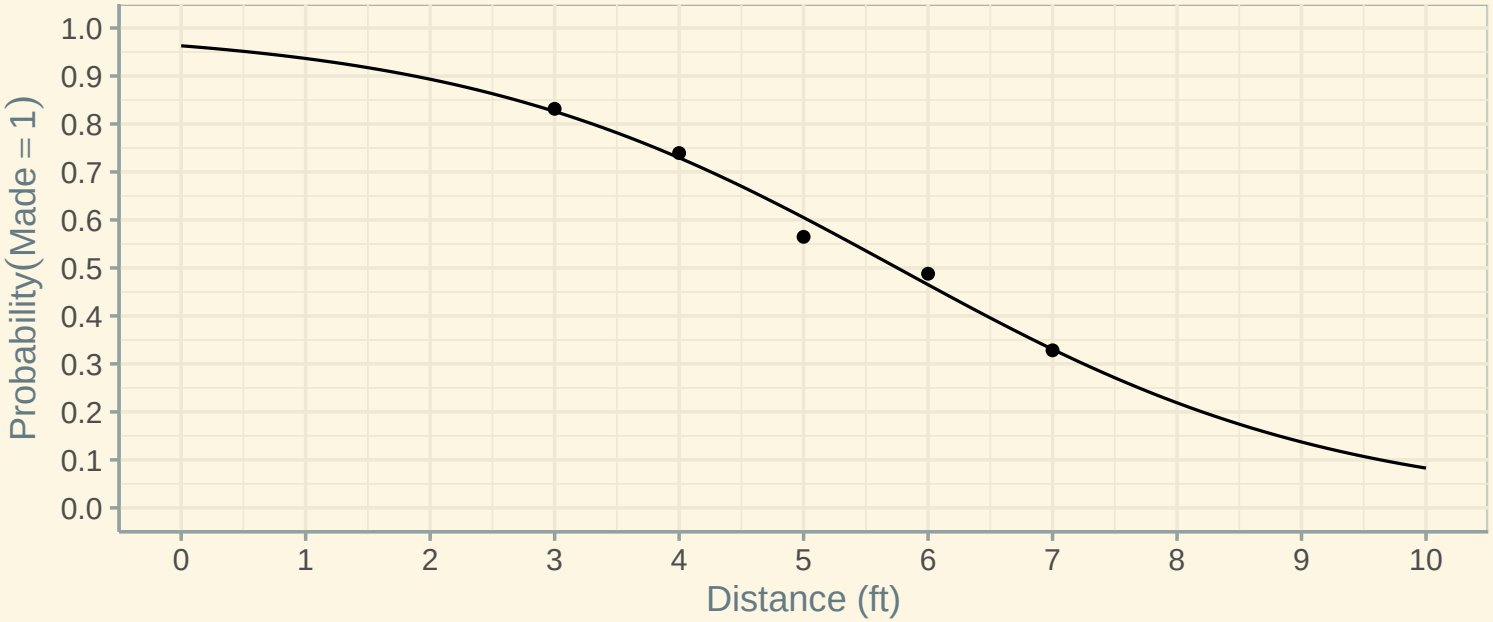
5. Find the sample odds ratio for success for 4 ft. vs. 3 ft; 5 vs. 4; 6 vs. 5; 7 vs. 6
6. Take the log of each of these to get the (additive) change in the log odds (logit). Should be slopes of lines "connecting the dots" (since $\Delta X = 1$).

Example: Golf Putts

Distance (ft)	3	4	5	6	7
# Made	84	88	61	61	44
# Missed	17	31	47	64	90
Odds	4.94	2.84	1.30	0.95	0.49
Log Odds	1.60	1.04	0.26	-0.05	-0.71
OR		0.575	0.457	0.734	0.513
Δ Log Odds		-0.56	-0.78	-0.31	-0.66

- In the data, successive ORs (changes in log odds) are different
- The model "smooths" the pattern

Example: Golf Putts (Probabilities)

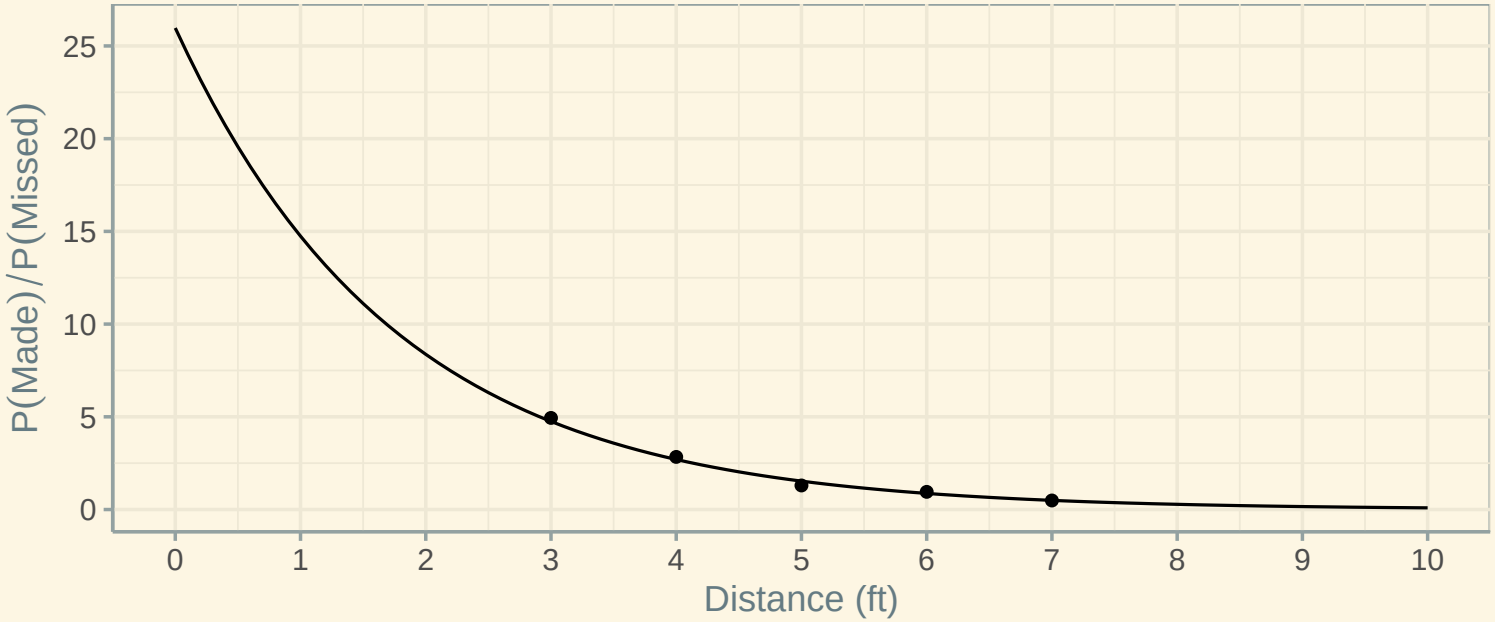



```
coef(golf_model) %>% round(digits = 2)

     (Intercept)    Distance
            3.26       -0.57
```

$$\text{Prob}(\texttt{Made} = 1) =: \hat{\pi} = \frac{e^{3.26 - 0.57 \cdot \texttt{Distance}}}{1 + e^{3.26 - 0.57 \cdot \texttt{Distance}}}$$

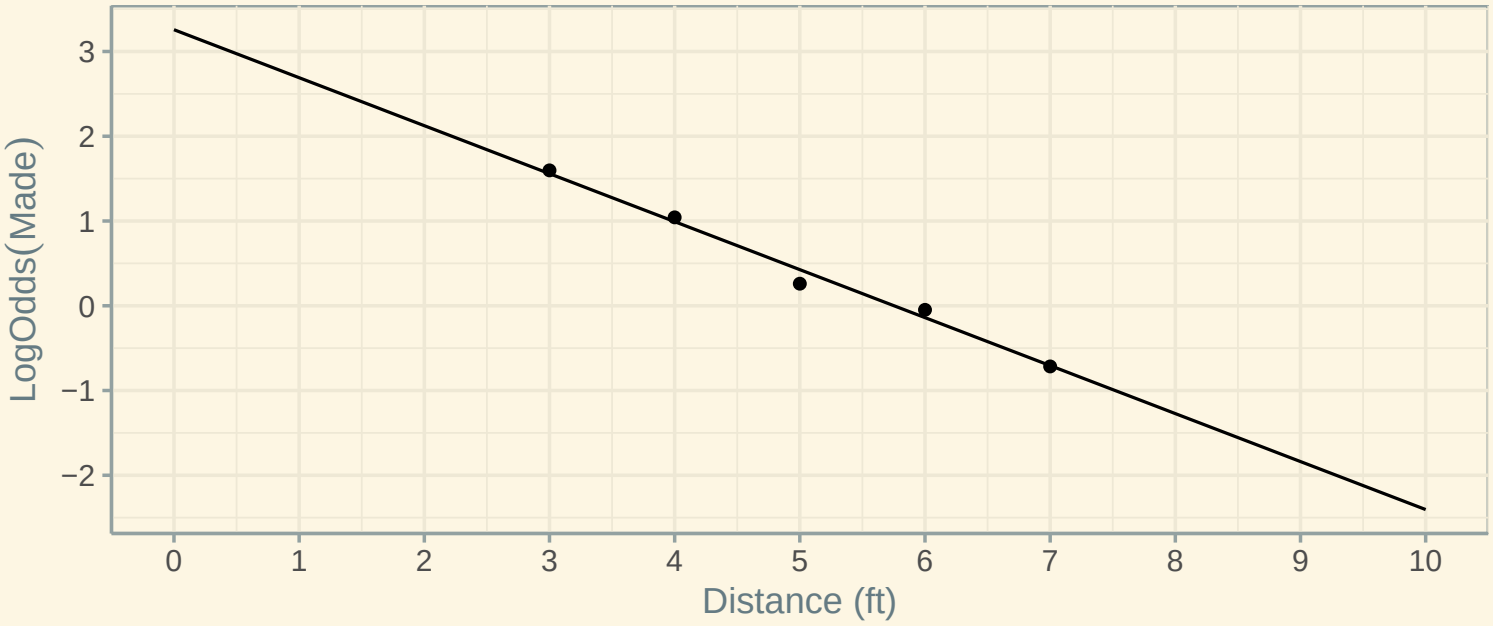
Example: Golf Putts (Odds)



$$\begin{aligned}\text{Odds}(\texttt{Made} = 1) &= \frac{\hat{\pi}}{1-\hat{\pi}} \\ &= e^{3.26 - 0.57 \cdot \texttt{Distance}} \\ &= e^{3.26}\left(e^{-0.57}\right)^{\texttt{Distance}}\end{aligned}$$

For each additional foot of distance the estimated **odds** of making the putt are **multiplied** by $e^{-0.57} = 0.57$

Example: Golf Putts (Log Odds)



$$\text{LogOdds}(\texttt{Made}) = 3.26 - 0.57 \cdot \texttt{Distance}$$

For each additional foot of distance, the estimated **log odds** of making the putt **decrease** by 0.57

Reconstructing Odds Ratio

- The logistic regression output from R gives us $\hat{\beta}_0$ and $\hat{\beta}_1$
- Unlike in linear regression, these are not very interpretable on their own.
- We have seen that β_1 corresponds to "rate of change in log odds"

$$\begin{aligned}\beta_1 &= \frac{\Delta \text{log odds}(Y=1)}{\text{one unit } \Delta \text{ in } X} \\ &= \text{log odds}(Y=1 \text{ at } x+1) - \text{log odds}(Y=1 \text{ at } x) \\ &= \log\left(\frac{\text{Odds}(Y=1 \text{ at } x+1)}{\text{Odds}(Y=1 \text{ at } x)}\right)\end{aligned}$$

- This is the **log odds ratio** associated with a one unit difference in X

The Odds Ratio

The **odds ratio** associated with a one unit difference in X is obtained by exponentiating the log odds ratio

$$e^{\beta_1} = \left(\frac{\text{Odds}(Y = 1 \text{ at } x + 1)}{\text{Odds}(Y = 1 \text{ at } x)} \right)$$

This implies that

$$\text{Odds}(Y = 1 \text{ at } x + 1) = e^{\beta_1} \text{Odds}(Y = 1 \text{ at } x)$$

More generally, for a difference of c units in X:

$$\text{Odds}(Y = 1 \text{ at } x + c) = \left(e^{\beta_1}\right)^c \text{Odds}(Y = 1 \text{ at } x)$$

Interpretation: for each 1 unit change in X, the **odds** that $Y = 1$ are **multiplied** by e^{β_1}