STAT 215
Logistic Regression I

Colin Reimer Dawson

Oberlin College

November 9-10, 2017

Outline

Quantitative Vs. Categorical Predictor and Response

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

Logistic Regression

Handout

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

Two Equivalent Forms of Logistic Regression

Probability Form

$$\pi = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

Logit Form

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

π : **Probability** that $Y = 1$

$\frac{\pi}{1-\pi}$: **Odds** that $Y = 1$

$\log\left(\frac{\pi}{1-\pi}\right)$: Log odds, or **logit** that $Y = 1$

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

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
4. Plot each of these against distance

Odds Ratios

Logit and Odds

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1 X}$$

- In the model, for each 1 unit increase in X, the logit increases by β_1.
- Equivalently: For each 1 unit increase in X, the odds are *multiplied* by e^{β_1}
- In other words, e^{β_1} is the *odds ratio* resulting from a one unit change in X, with β_1 the *log odds ratio*.

Odds Ratios

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

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

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 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 fits a constant ratio (slope for log odds)

7. Draw a single line through your logit plot and get an estimated slope and intercept. These are your $\hat{\beta}_0$ and $\hat{\beta}_1$.

Example: Golf Putts

```
library("mosaic")
Putts <- data.frame(Distance = 3:7, Made = c(84,88,61,61,44),
                    Total = c(101,119,108,125,134))
Putts <- mutate(Putts, PropMade = Made / Total)
model <- glm(PropMade ~ Distance, weights = Total,
             data = Putts, family = "binomial")
model
```

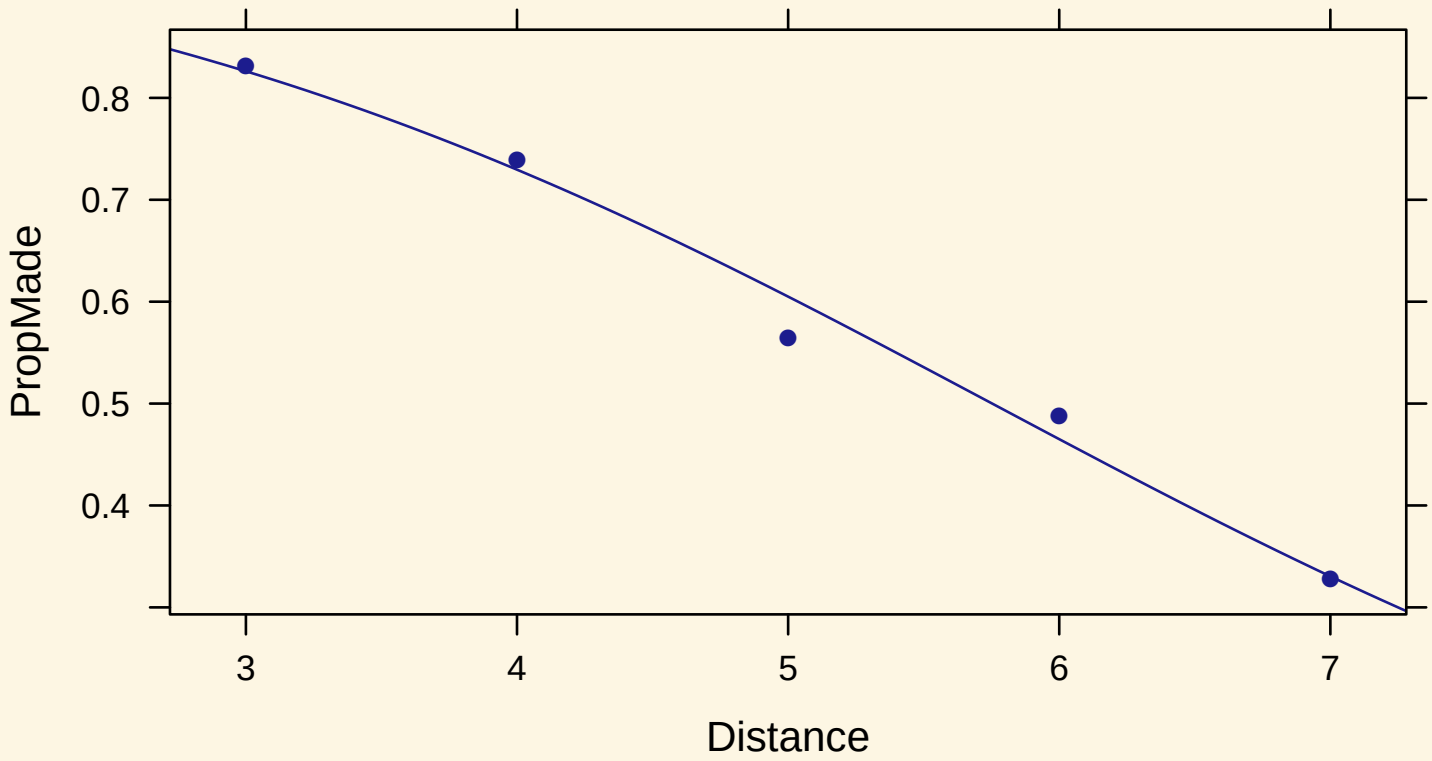
```
Call:  glm(formula = PropMade ~ Distance, family = "binomial", data = Putt
    weights = Total)

Coefficients:
(Intercept)     Distance
     3.2568      -0.5661

Degrees of Freedom: 4 Total (i.e. Null);  3 Residual
Null Deviance:      81.39
Residual Deviance: 1.069  AIC: 30.18
```

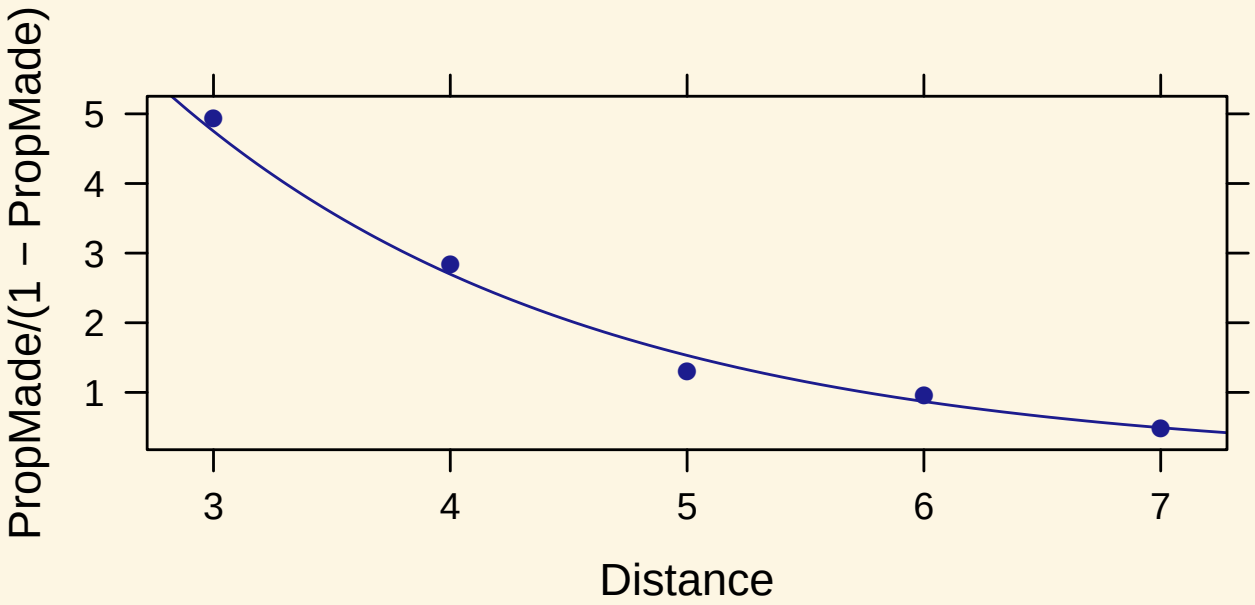
Example: Golf Putts (Probabilities)

```
xyplot(PropMade ~ Distance, data = Putts)
prob.hat <- makeFun(model)
plotFun(prob.hat(Distance) ~ Distance, add = TRUE)
```



Example: Golf Putts (Odds)

```
odds.hat <- makeFun(model, transformation = function(x){x/(1-x)})
xyplot(PropMade/(1 - PropMade) ~ Distance, data = Putts)
plotFun(odds.hat(Distance) ~ Distance, add = TRUE)
```

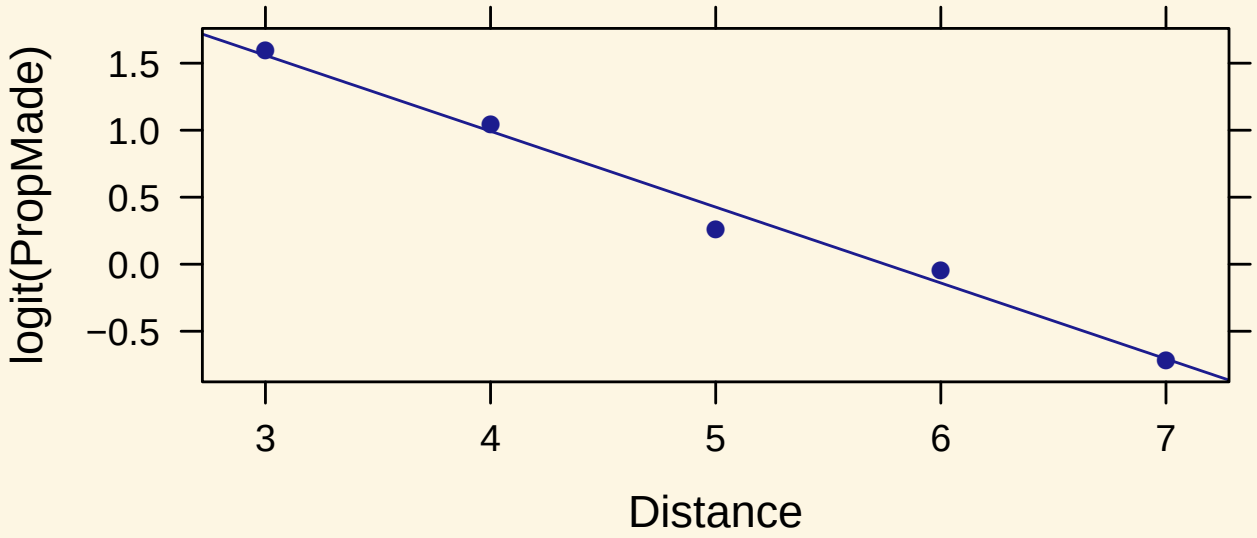



```
exp(-0.5661) ## Odds ratio for a one foot increase in Distance

     [1] 0.5677353
```

Example: Golf Putts (Log Odds)

```
log.odds.hat <- makeFun(model, transformation = logit)
xyplot(logit(PropMade) ~ Distance, data = Putts)
plotFun(log.odds.hat(Distance) ~ Distance, add = TRUE)
```




```
-0.5661 ## Log (odds ratio) / rate of change in log odds / slope of logit
```

Reconstructing Odds Ratio

- The logistic regression output from R gives us $\hat{\beta}_0$ and $\hat{\beta}_1$. But 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". (Slightly) better to convert to "odds ratio" per unit change in X.
- What do we do to β_1 to get this?

Choosing $\hat{\beta}_0$ and $\hat{\beta}_1$

Recall that in linear regression, we choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize

$$RSS = \sum_i (Y_i - f(X_i))^2 = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X)^2$$

For a logistic model, choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to *maximize* **the probability of the data according to the model**.

$$Pr(\text{Data}|\text{Model}) = \prod_{i=1}^{n} \hat{\pi}_i^{Y_i} (1 - \hat{\pi}_i)^{1-Y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_i}} \right)^{Y_i} \left(\frac{1}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_i}} \right)^{1-Y_i}$$

Maximum Likelihood

- $Pr(\text{Data}|\text{Model})$ is called the **likelihood** of the model.
- In fact, when we assume heteroskedastic Normal residuals, the RSS is the negative log likelihood.
- So we've secretly been doing max likelihood this whole time.
- But whereas MLE for Normal-linear model was a calculus problem, MLE for logistic requires an iterative algorithm.