

Binary outcomes

Jeffrey Leek, Assistant Professor of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Key ideas

- Frequently we care about outcomes that have two values
 - Alive/dead
 - Win/loss
 - Success/Failure
 - etc
- Called binary outcomes or 0/1 outcomes
- Linear regression (like we've seen) may not be the best

Example: Baltimore Ravens

Baltimore Ravens Sign in to personalize
AFC North

Clubhouse Stats Schedule Roster Splits Depth Chart Transactions Rankings Photos Stadium News Forum

Sun Feb 3 Sun Feb 3

@ **W** 34-31
San Francisco 49ers

Pass: Kaepernick 302 yds
Rush: Gore 110 yds
Rec: Crabtree 109 yds

Recap »
Box Score »

Final Superbowl

@

Baltimore (10-6) **San Francisco** (11-4-1)

	1	2	3	4	T
BAL	7	14	7	6	34
SF	3	3	17	8	31

Recap » Box Score »

2012 Season

Record:
Overall: 10-6
vs AFC North: 4-2
vs AFC: 8-4

Team leaders:
Pass: Flacco 3817 yds
Rush: Rice 1143 yds
Rec: Boldin 921 yds

BALTIMORE TEAMS

2012 OVERALL NFL RANKINGS

PASSING YDS	RUSHING YDS	OPP PASSING YDS	OPP RUSHING YDS
4,111	2,007	4,411	1,100
4,111	2,007	4,411	1,100
4,111	2,007	4,411	1,100

Tickets Shop

http://espn.go.com/nfl/team/_/name/bal/baltimore-ravens

Ravens Data

```
download.file("https://dl.dropbox.com/u/7710864/data/ravensData.rda",  
             destfile="./data/ravensData.rda",method="curl")  
load("./data/ravensData.rda")  
head(ravensData)
```

	ravenWinNum	ravenWin	ravenScore	opponentScore
1	1	W	24	9
2	1	W	38	35
3	1	W	28	13
4	1	W	34	31
5	1	W	44	13
6	0	L	23	24

Linear regression

$$RW_i = b_0 + b_1 RS_i + e_i$$

RW_i - 1 if a Ravens win, 0 if not

RS_i - Number of points Ravens scored

b_0 - probability of a Ravens win if they score 0 points

b_1 - increase in probability of a Ravens win for each additional point

e_i - variation due to everything we didn't measure

Linear regression in R

```
lmRavens <- lm(ravensData$ravenWinNum ~ ravensData$ravenScore)
summary(lmRavens)
```

Call:

```
lm(formula = ravensData$ravenWinNum ~ ravensData$ravenScore)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.730	-0.508	0.182	0.322	0.572

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.28503	0.25664	1.11	0.281
ravensData\$ravenScore	0.01590	0.00906	1.76	0.096 .

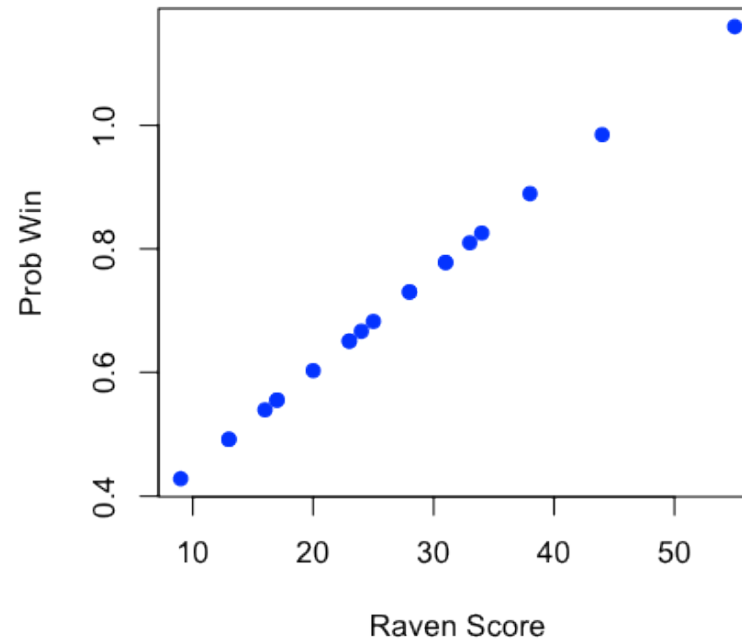
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.446 on 18 degrees of freedom

Multiple R-squared: 0.146, Adjusted R-squared: 0.0987

Linear regression

```
plot(ravensData$ravenScore, lmRavens$fitted, pch=19, col="blue", ylab="Prob Win", xlab="Raven Score")
```



Odds

Binary Outcome 0/1

$$RW_i$$

Probability (0,1)

$$\Pr(RW_i | RS_i, b_0, b_1)$$

Odds $(0, \infty)$

$$\frac{\Pr(RW_i | RS_i, b_0, b_1)}{1 - \Pr(RW_i | RS_i, b_0, b_1)}$$

Log odds $(-\infty, \infty)$

$$\log \left(\frac{\Pr(RW_i | RS_i, b_0, b_1)}{1 - \Pr(RW_i | RS_i, b_0, b_1)} \right)$$

Linear vs. logistic regression

Linear

$$RW_i = b_0 + b_1 RS_i + e_i$$

or

$$E[RW_i | RS_i, b_0, b_1] = b_0 + b_1 RS_i$$

Logistic

$$\Pr(RW_i | RS_i, b_0, b_1) = \frac{\exp(b_0 + b_1 RS_i)}{1 + \exp(b_0 + b_1 RS_i)}$$

or

$$\log\left(\frac{\Pr(RW_i | RS_i, b_0, b_1)}{1 - \Pr(RW_i | RS_i, b_0, b_1)}\right) = b_0 + b_1 RS_i$$

Interpreting Logistic Regression

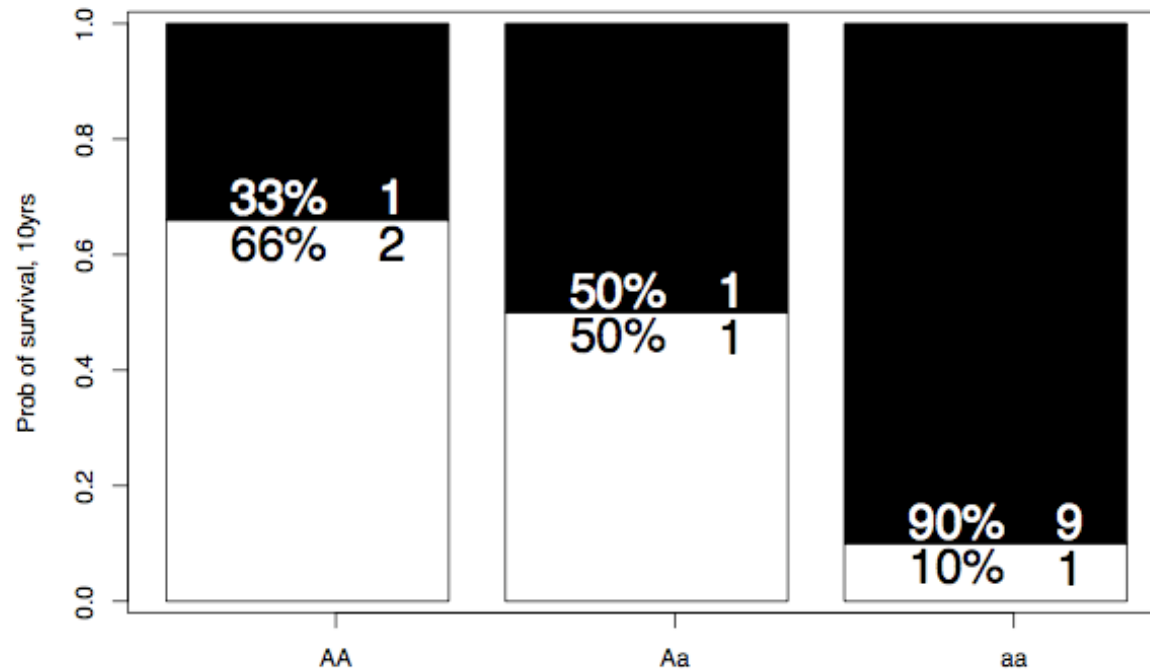
$$\log\left(\frac{\Pr(RW_i|RS_i, b_0, b_1)}{1 - \Pr(RW_i|RS_i, b_0, b_1)}\right) = b_0 + b_1 RS_i$$

b_0 - Log odds of a Ravens win if they score zero points

b_1 - Log odds ratio of win probability for each point scored (compared to zero points)

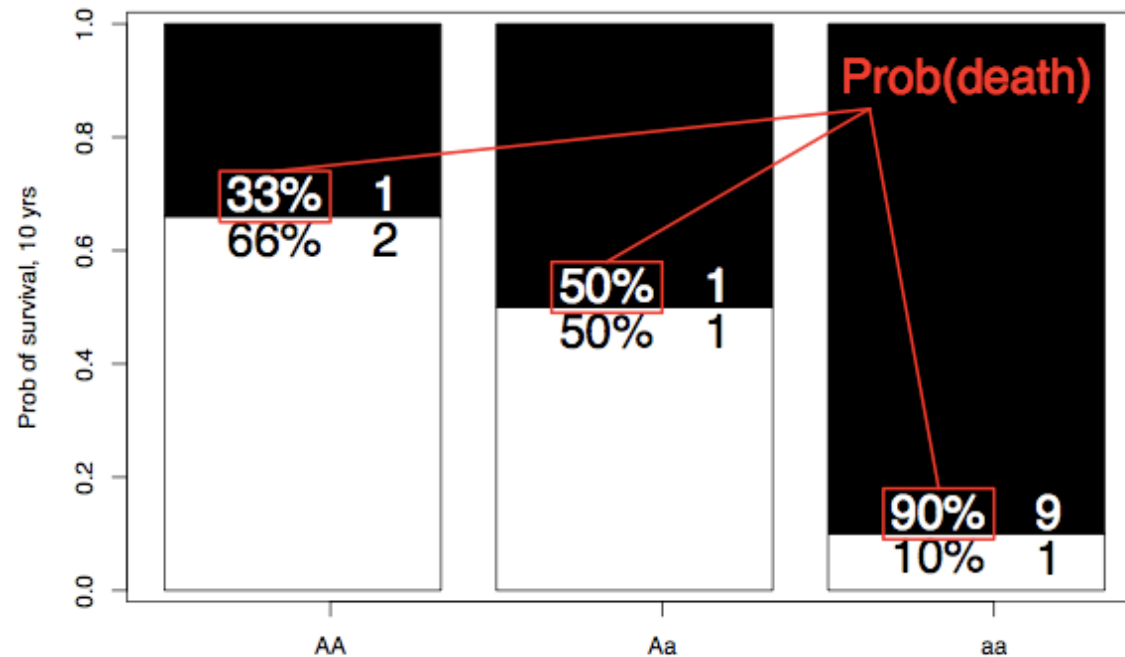
$\exp(b_1)$ - Odds ratio of win probability for each point scored (compared to zero points)

Explaining Odds



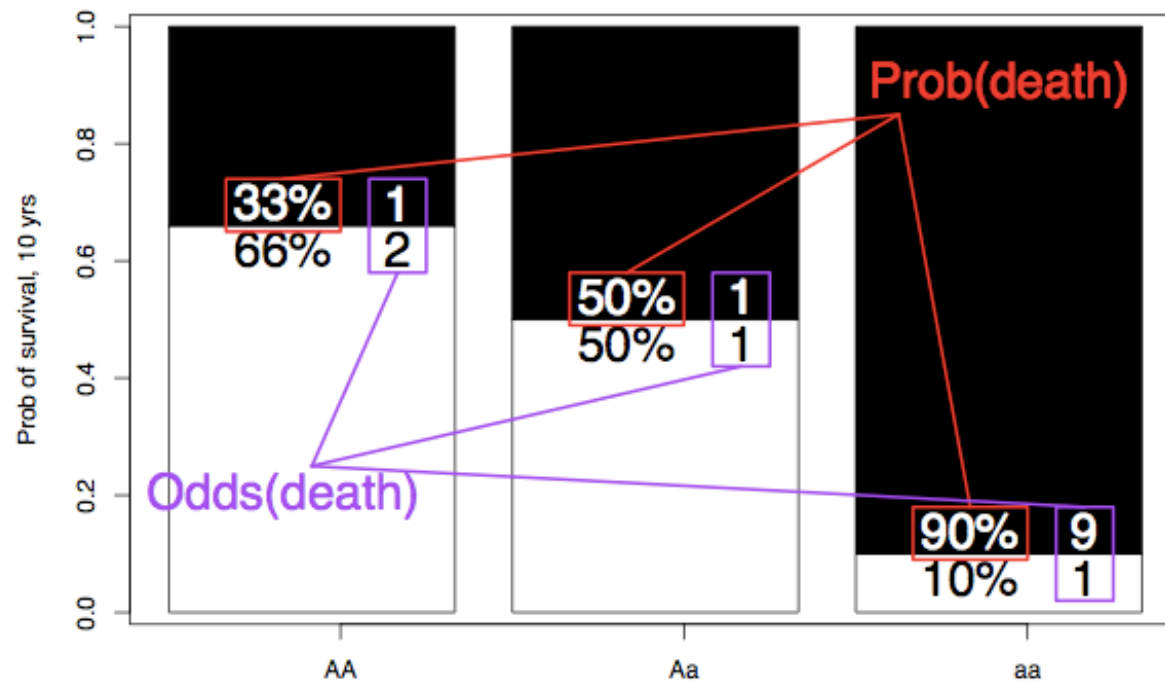
[via Ken Rice](#)

Probability of Death



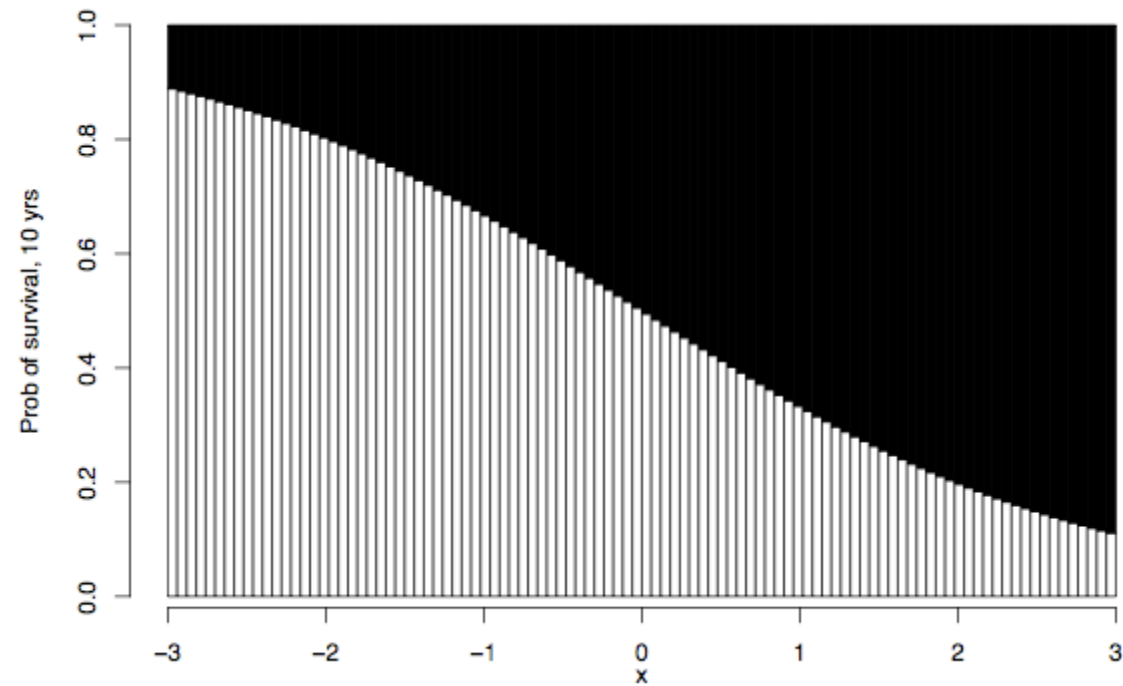
[via Ken Rice](#)

Odds of Death



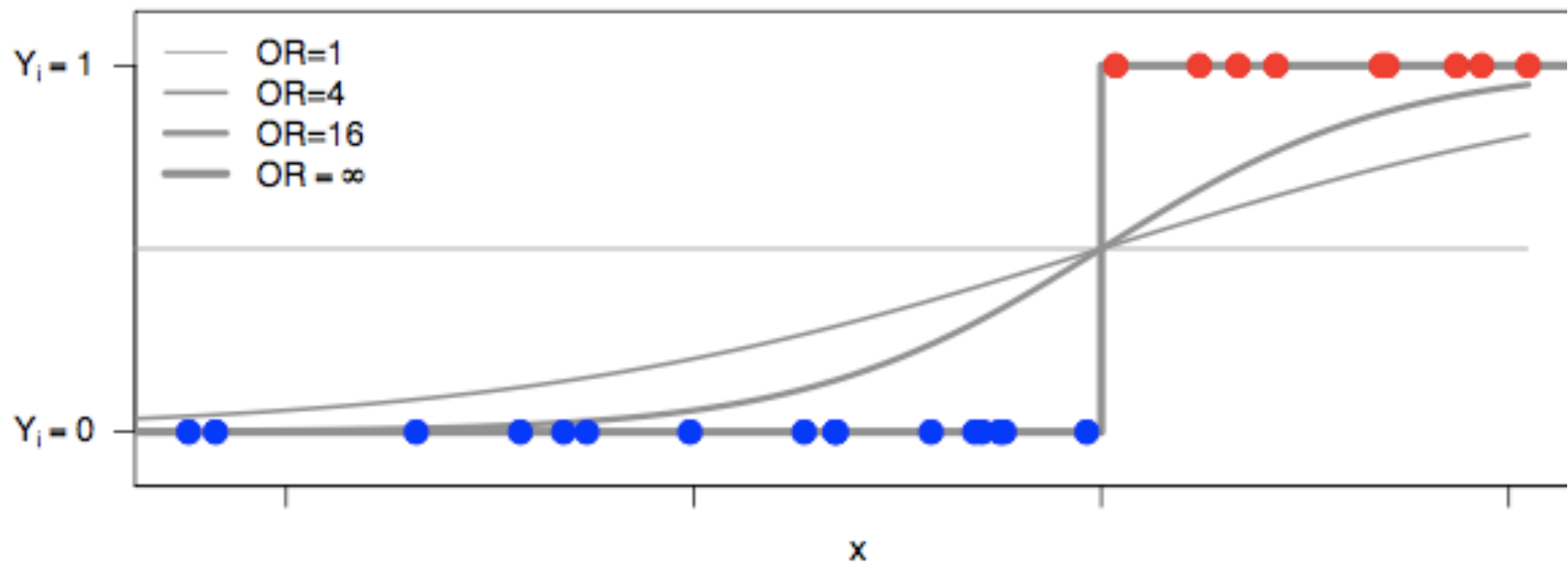
[via Ken Rice](#)

Odds Ratio = 1, Continuous Covariate



[via Ken Rice](#)

Different odds ratios



[via Ken Rice](#)

Ravens logistic regression

```
logRegRavens <- glm(ravensData$ravenWinNum ~ ravensData$ravenScore,family="binomial")
summary(logRegRavens)
```

Call:

```
glm(formula = ravensData$ravenWinNum ~ ravensData$ravenScore,
     family = "binomial")
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.758	-1.100	0.530	0.806	1.495

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.6800	1.5541	-1.08	0.28
ravensData\$ravenScore	0.1066	0.0667	1.60	0.11

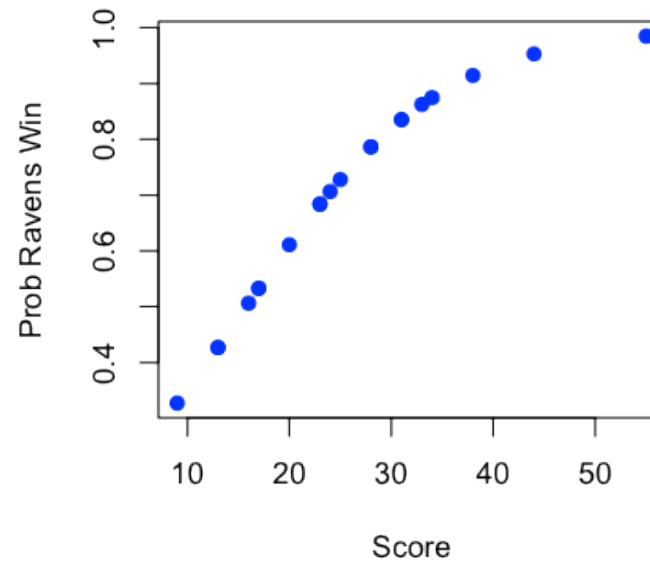
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 24.435 on 19 degrees of freedom

16/22

Ravens fitted values

```
plot(ravensData$ravenScore, logRegRavens$fitted, pch=19, col="blue", xlab="Score", ylab="Prob Ravens Win")
```



Odds ratios and confidence intervals

```
exp(logRegRavens$coeff)
```

```
(Intercept) ravensData$ravenScore
0.1864      1.1125
```

```
exp(confint(logRegRavens))
```

```

                2.5 % 97.5 %
(Intercept)    0.005675  3.106
ravensData$ravenScore 0.996230 1.303
```

ANOVA for logistic regression

```
anova(logRegRavens, test="Chisq")
```

Analysis of Deviance Table

Model: binomial, link: logit

Response: ravensData\$ravenWinNum

Terms added sequentially (first to last)

	Df	Deviance	Resid.	Df	Resid.	Dev	Pr(>Chi)
NULL			19		24.4		
ravensData\$ravenScore	1	3.54	18		20.9	0.06	.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Simpson's paradox

	Treatment A	Treatment B
Small Stones	Group 1 93% (81/87)	Group 2 87% (234/270)
Large Stones	Group 3 73% (192/263)	Group 4 69% (55/80)
Both	78% (273/350)	83% (289/350)

http://en.wikipedia.org/wiki/Simpson's_paradox

Interpreting Odds Ratios

- Not probabilities
- Odds ratio of 1 = no difference in odds
- Log odds ratio of 0 = no difference in odds
- Odds ratio < 0.5 or > 2 commonly a "moderate effect"
- Relative risk $\frac{\Pr(RW_i | RS_i=10)}{\Pr(RW_i | RS_i=0)}$ often easier to interpret, harder to estimate
- For small probabilities $RR \approx OR$ but **they are not the same!**

[Wikipedia on Odds Ratio](#)

Further resources

- [Wikipedia on Logistic Regression](#)
- [Logistic regression and glms in R](#)
- Brian Caffo's lecture notes on: [Simpson's paradox](#), [Case-control studies](#)
- [Open Intro Chapter on Logistic Regression](#)