Visualizing Multiple Linear Regression and Binary Logistic Regression Models

Skill Building Workshop

Evaluation 2014, Denver, CO

Before

Multiple linear regression

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-13.667	2.647		-5.164	.000
	Age Age, in years	.702	.044	.323	15.961	.000
	Weight Weight, in kg	.906	.049	.717	18.490	.000
	BSA Body Surface Area	4.627	1.521	.116	3.042	.008

a. Dependent Variable: BP Blood Pressure

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.997[a]	.995	.994	.437

a. Predictors: (Constant), BSA Body Surface Area, Age Age, in years, Weight Weight, in kg

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	556.944	3	185.648	971.934	.000[a]
	Residual	3.056	16	.191		
	Total	560.000	19			

a. Predictors: (Constant), BSA Body Surface Area, Age Age, in years, Weight Weight, in kg
b. Dependent Variable: BP Blood Pressure

Binary logistic regression

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	pclass			120.536	2	.000			
	pclass(1)	1.557	.143	117.934	1	.000	4.743	3.581	6.282
	pclass(2)	.787	.149	27.970	1	.000	2.197	1.641	2.941
	Constant	-1.071	.086	154.497	1	.000	.343		

a. Variable(s) entered on step 1: pclass.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1613.259[a]	.093	.126

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Classification Table[a]

Observed			Predicted survived 0	Predicted survived 1	Predicted Percentage Correct
Step 1	survived	0	686	123	84.8
		1	300	200	40.0
	Overall Percentage				67.7

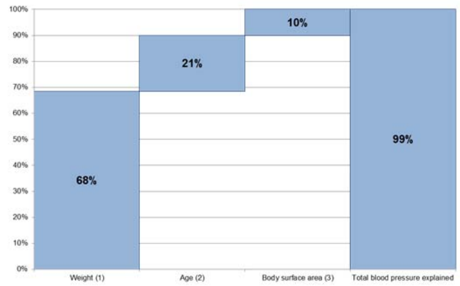
a. The cut value is .500

2

After

Multiple linear regression

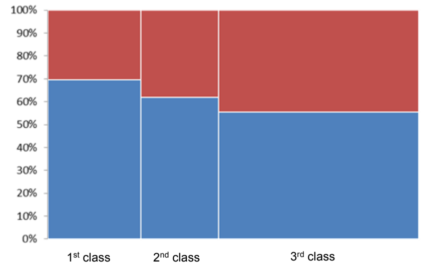
Figure 1. Blood Pressure Drivers



This chart presents Pratt Index scores that express multiple linear regression coefficients, where the dependent variable is blood pressure, and three independent variables, as a percentage of total variance explained by the model (standardized equation: $y = .717x_1 + .323x_2 + .116x_3$; R^2=.99, all coefficients are statistically significant, $a < 0.01$).

Binary logistic regression

Figure 2. Passenger Class as Survival Driver



This chart presents conditional probabilities and odds that express binary logistic regression coefficients, where the dependent variable is survival, the model equation: $\ln(y) = -1.071 + 1.557x_1 + .787_2$; Correct predictions: 68%. Omnibus test of model coefficients: χ^2=127.8, all coefficients are statistically significant, $a < 0.01$. -2Log Likelihood = 1,613; Cox & Snell R^2 = .093; Nagelkerke R^2 = .126.

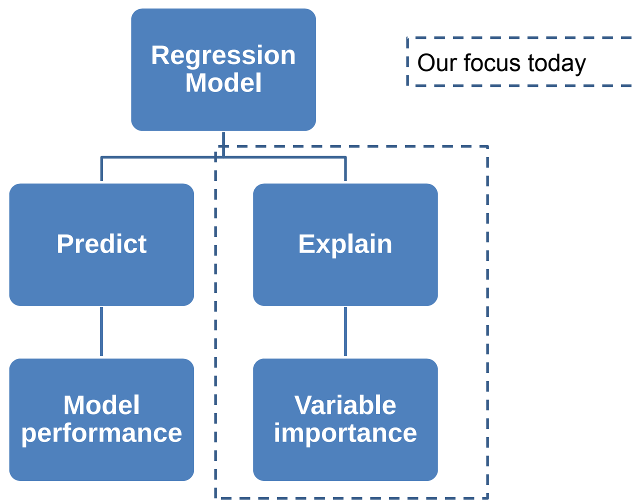
3

In this session

- **Introduction**
- **Linear regression**
  - Exercise 1: Calculate Pratt Index
- **Mosaic plots**
  - Exercise 2: Build a simple mosaic plot
- **Logistic regression**
  - Exercise 3: Build a mosaic plot for a binary model

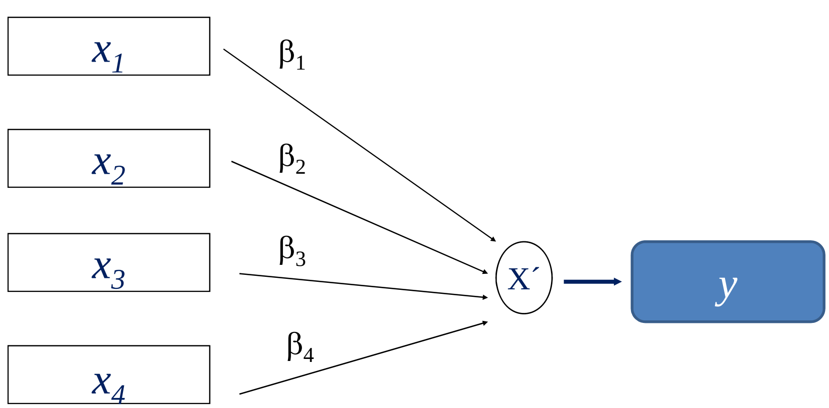
peacecorps.gov

Introduction: Model Goals



5

Introduction: Multiple Regression



6

Introduction: the Math

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_j x_j + \beta_q x_q + \varepsilon$$

$\alpha =$ Constant or intercept

$\beta_1 \rightarrow \beta_q =$ Coefficients

$x_1 \rightarrow x_q =$ Explanatory variables

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_j x_j + \beta_q x_q + \varepsilon$$

$p =$ Probability of event occurring

$\frac{p}{1-p} =$ Odds ratio

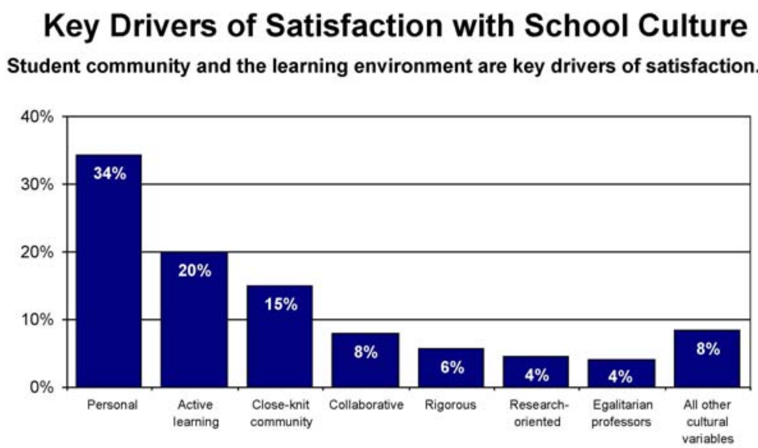
7

In this session

- Introduction
- **Linear regression**
  - Exercise 1: Calculate Pratt Index
- Mosaic plots
  - Exercise 2: Build a simple mosaic plot
- Logistic regression
  - Exercise 3: Build a mosaic plot for a binary model

peacecorps.gov

Pratt Index—Example 1



Source: GMAC, Impact of School Culture: European Full-Time MBA Programs, 2008

9

Pratt Index—Partition of R^2

Beta coefficient of a variable x_j from the standardized regression equation

Simple correlation between a variable x_j and dependent variable y

$$d_j = \frac{\beta_j r_j}{R^2}$$

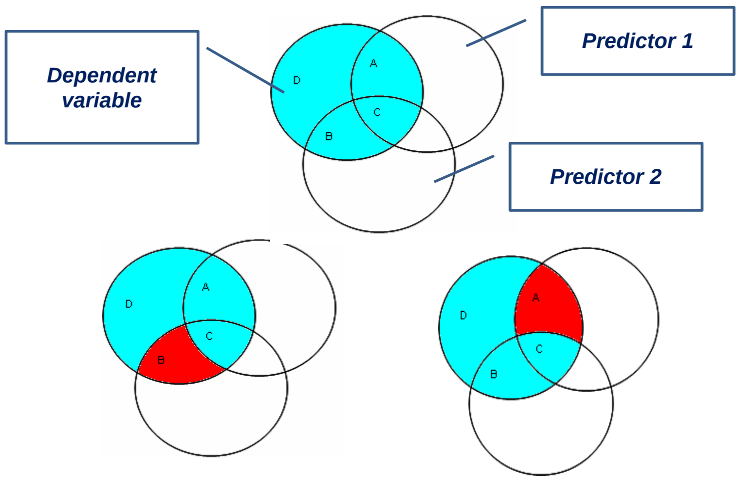
Proportion of variance R^2 accounted for by a variable x_j

Variation in the values of y explained by X

Just for reference: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_j x_j + \beta_p x_p + \varepsilon_i$

10

Closest to Semipartial Correlation



B/(A+B+C+D)

A/(A+B+C+D)

Blood Pressure Dataset

	A	B	C	D	E	F	G	H
1	Pt	BP	Age	Weight	BSA	Dur	Pulse	Stress
2	1	105	47	85.4	1.75	5.1	63	33
3	2	115	49	94.2	2.10	3.8	70	14
4	3	116	49	95.3	1.98	8.2	72	10
5	4	117	50	94.7	2.01	5.8	73	99
6	5	112	51	89.4	1.89	7.0	72	95
7	6	121	48	99.5	2.25	9.3	71	10
8	7	121	49	99.8	2.25	2.5	69	42
9	8	110	47	90.9	1.90	6.2	66	8
10	9	110	49	89.2	1.83	7.1	69	62
11	10	114	48	92.7	2.07	5.6	64	35
12	11	114	47	94.4	2.07	5.3	74	90
13	12	115	49	94.1	1.98	5.6	71	21
14	13	114	50	91.6	2.05	10.2	68	47
15	14	106	45	87.1	1.92	5.6	67	80
16	15	125	52	101.3	2.19	10.0	76	98
17	16	114	46	94.5	1.98	7.4	69	95
18	17	106	46	87.0	1.87	3.6	62	18
19	18	113	46	94.5	1.90	4.3	70	12
20	19	110	48	90.5	1.88	9.0	71	99
21	20	122	56	95.7	2.09	7.0	75	99

Regression Output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.997[a]	.995	.994	.437

a. Predictors: (Constant), BSA Body Surface Area, Age Age, in years, Weight Weight, in kg

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	556.944	3	185.648	971.934	.000[a]
	Residual	3.056	16	.191		
	Total	560.000	19			

a. Predictors: (Constant), BSA Body Surface Area, Age Age, in years, Weight Weight, in kg
b. Dependent Variable: BP Blood Pressure

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	-13.667	2.647		-5.164	.000			
	Age Age, in years	.702	.044	.323	15.961	.000	.659	.970	.295
	Weight Weight, in kg	.906	.049	.717	18.490	.000	.950	.977	.341
	BSA Body Surface Area	4.627	1.521	.116	3.042	.008	.866	.605	.056

a. Dependent Variable: BP Blood Pressure

Correlations

		BP Blood Pressure	Age Age, in years	Weight Weight, in kg	BSA Body Surface Area
BP Blood Pressure	Pearson Correlation	1	.659**	.950**	.866**
	Sig. (2-tailed)		.002	.000	.000
	N	20	20	20	20

13

Pratt Index—How To

Variable	β	r	β*r	%
Age in years	0.323	0.659	0.213	21.4%
Weight in kg	0.717	0.950	0.681	68.5%
Body Surface Area (BSA)	0.116	0.866	0.101	10.1%
SUM			0.995	100.0%

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.997[a]	.995	.994	.437

a. Predictors: (Constant), BSA Body Surface Area, Age Age, in years, Weight Weight, in kg

14

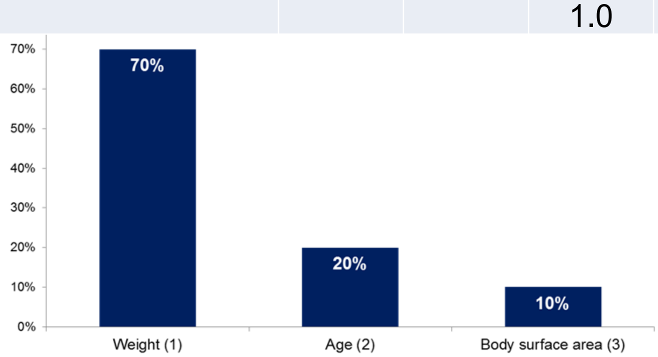
Pratt Index Exercise

1. **Multiply β by corresponding r**
   - Round inputs to one decimal point
   - Round products to one decimal point
2. **Add the products to ensure that the sum is equal to R^2**
   - Round to 1
3. **Divide each product by R^2**
   - Express as a percentage

peacecorps.gov

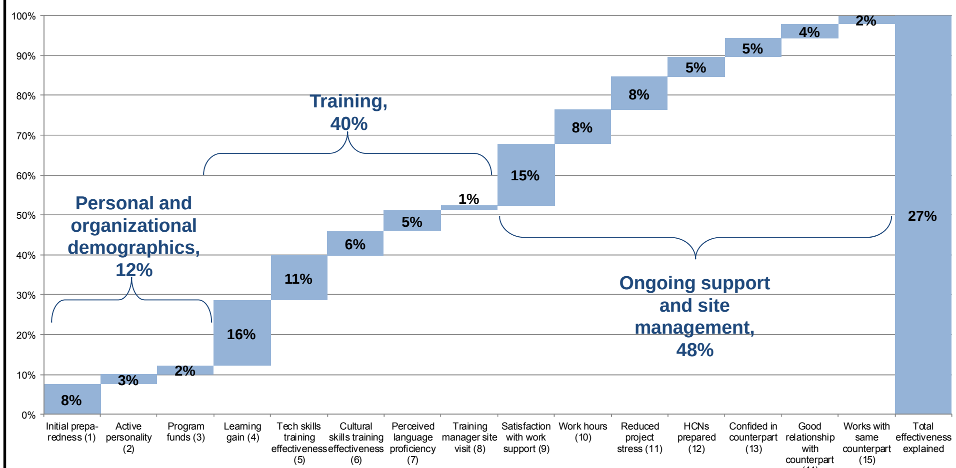
Pratt Index—Check Your Worksheet

Variable	β	r	β*r	%
Age in years	0.3	0.7	0.2	20%
Weight in kg	0.7	1.0	0.7	70%
Body Surface Area (BSA)	0.1	0.9	0.1	10%
SUM			1.0	100%



16

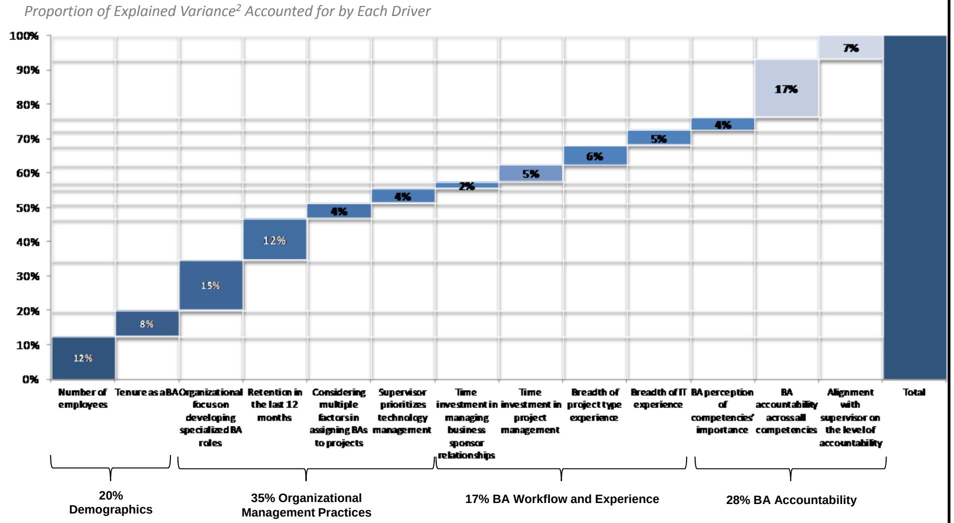
Pratt Index—Example 2
Becoming More Effective Volunteer



This chart presents Pratt Index scores that express multiple linear regression coefficients, where the dependent variable is the Volunteer effectiveness measure, and 15 independent variables covering Volunteer personal and post organizational characteristics, training, and ongoing staff support and site management, as a percentage of total variance explained by the model (standardized equation: $y = .184x_1 + .058x_2 + .075x_3 + .220x_4 + .090x_5 + .052x_6 + .075x_7 + .040x_8 + .116x_9 + .113x_{10} + .089x_{11} + .053x_{12} + .074x_{13} + .048x_{14} + .042x_{15}$; $R^2=.27$, Multiple R=.52, all coefficients are statistically significant, $a < 0.01$).
Source: Peace Corps, Voice of the Volunteer, 2013

Pratt Index—Example 3

Relative Importance of BA Effectiveness Drivers[1]
Proportion of Explained Variance[2] *Accounted for by Each Driver*



1. Pratt Index.
2. Multiple linear regression model, $R^2=0.44$
Source: CEB, BA Effectiveness Diagnostic. 2012

18

Assumptions

1) Relative importance depends only on the means, variances and correlations of y, x_1, x_2, …, x_j, x_p.
2) Relative importance is not affected by linear transformations of any variable.
3) The relative importance of x_1 to x_2 is as m to n => positive $\beta_j r_j$!
4) The non-singular linear transformation of a subset of (x_1, …, x_q) into the subset (x_1', …, x_q') does not affect its importance relative to other variables.
5) The addition of a pure noise variable, independent of y and x_1, …, x_p, to a subset of variables does not affect importance of the subset relative to other variables.

19

Major Criticisms

- Negative $\beta_j r_j$ = negative importance?
- x orthogonal to y, but nonetheless increases R^2.

20

In this session

- Introduction
- Linear regression
  - Exercise 1: Calculate Pratt Index
- **Mosaic plots**
  - Exercise 2: Build a simple mosaic plot
- Logistic regression
  - Exercise 3: Build a mosaic plot for a binary model

peacecorps.gov

Titanic Dataset

row.names	pclass	survived	name	age	embarked	home.dest	room	ticket	boat	sex
1	1st	1	Allen, Mis	29	Southamp	St Louis, M	B-5	24160 L221	2	female
2	1st	0	Allison, M	2	Southamp	Montreal,	C26			female
3	1st	0	Allison, M	30	Southamp	Montreal,	C26		-135	male
4	1st	0	Allison, M	25	Southamp	Montreal,	C26			female
5	1st	1	Allison, M	0.9167	Southamp	Montreal,	C22		11	male
6	1st	1	Anderson,	47	Southamp	New York,	E-12		3	male
7	1st	1	Andrews,	63	Southamp	Hudson, N	D-7	13502 L77	10	female
8	1st	0	Andrews,	39	Southamp	Belfast, N	A-36			male
9	1st	1	Appleton,	58	Southamp	Bayside, C	C-101		2	female
10	1st	0	Artagavey	71	Cherbourg	Montevideo, Uruguay			-22	male
11	1st	0	Astor, Col	47	Cherbourg	New York, NY		17754 L224	-124	male
12	1st	1	Astor, Mrs	19	Cherbourg	New York, NY		17754 L224	4	female
13	1st	1	Aubert, M	NA	Cherbourg	Paris, Frar	B-35	17477 L69	9	female
14	1st	1	Barkworth	NA	Southamp	Hessle, Yo	A-23		B	male
15	1st	0	Baumann,	NA	Southamp	New York, NY				male
16	1st	1	Baxter, Mr	50	Cherbourg	Montreal,	B-58/60		6	female
17	1st	0	Baxter, Mr	24	Cherbourg	Montreal,	B-58/60			male
18	1st	0	Beattie, M	36	Cherbourg	Winnipeg	C-6			male
19	1st	1	Beckwith,	37	Southamp	New York,	D-35		5	male
20	1st	1	Beckwith,	47	Southamp	New York,	D-35		5	female
21	1st	1	Behr, Mr K	26	Cherbourg	New York,	C-148		5	male
1310	3rd	0	Zakarian, I	NA						male
1311	3rd	0	Zenn, Mr P	NA						male
1312	3rd	0	Zievens, R	NA						female
1313	3rd	0	Zimmerm	NA						male

22

What is a Mosaic Plot?

- A tool to display relationships among multiple categorical variables

3 x 2 x 2	Did not survive	Did not survive	Survived	Survived
	Male	Female	Male	Female
1st class	120	9	59	134
2nd class	148	13	25	94
3rd class	440	134	58	79

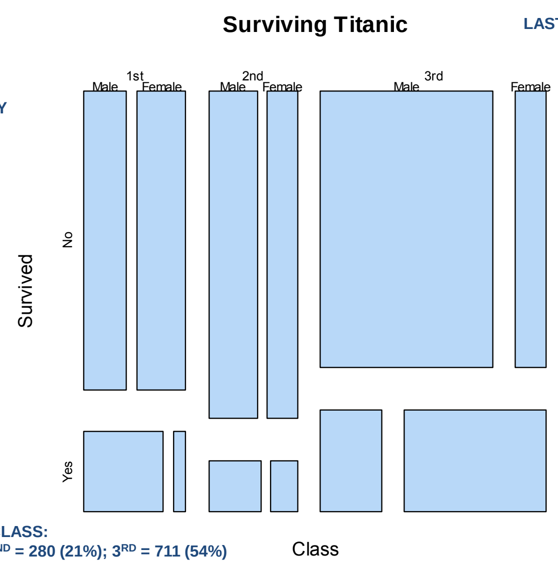
TOTAL = 1,313
1st = 322 => MALE = 179; FEMALE = 143
2nd = 280 => MALE = 173; FEMALE = 107
3rd = 711 => MALE = 498; FEMALE = 213

23

Visualizing 3 x 2 x 2

LAST SPLIT BY GENDER

SECOND SPLIT BY SURVIVAL



FIRST SPLIT BY CLASS:
1ST = 322 (25%); 2ND = 280 (21%); 3RD = 711 (54%)

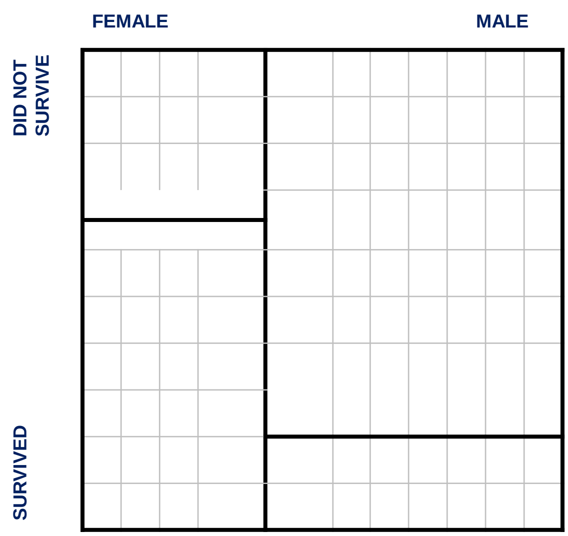
24

Mosaic Plot Exercise

1. **Square with a length of one**
   – Divide the square vertically
2. **First split: treatment**
   – Divide the square vertically
   – 35% female; 65% male
3. **Last split: outcome**
   – Divide the square horizontally
   – Females: 67% survived; 33% died
   – Males: 17% survived; 83% died

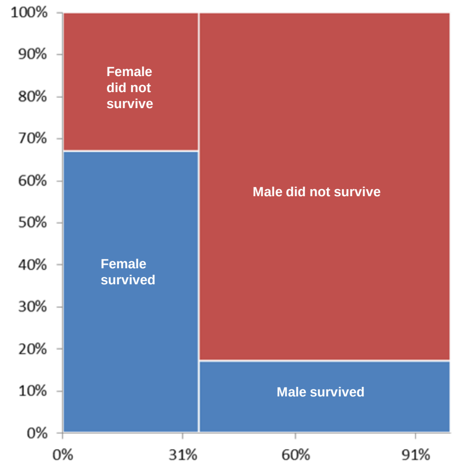
peacecorps.gov

Mosaic Plot—Check Your Worksheet



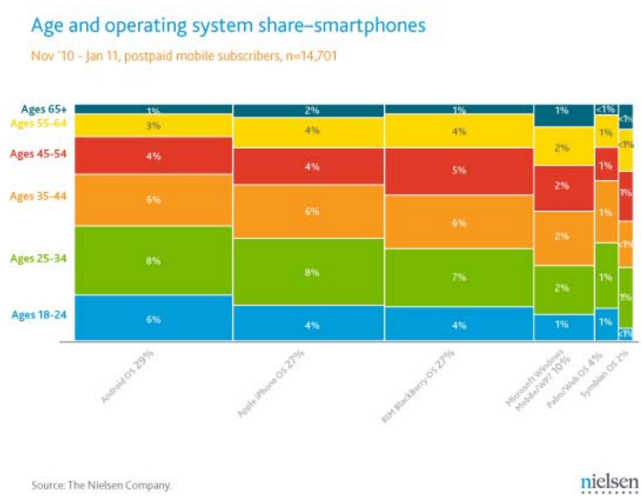
26

Mosaic Plot—Example 1



27

Mosaic Plot—Example 2



Source: The Nielsen Company, 2011

28

In this session

- Introduction
- Linear regression
  - Exercise 1: Calculate Pratt Index
- Mosaic plots
  - Exercise 2: Build a simple mosaic plot
- **Logistic regression**
  - Exercise 3: Build a mosaic plot for a binary model

peacecorps.gov

Regression Output

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1358.722[a]	.221	.306

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Classification Table[a]

Observed			Predicted: survived 0	Predicted: survived 1	Percentage Correct
Step 1	survived	0	708	156	81.9
		1	142	307	68.4
	Overall Percentage				77.3

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	sex	2.284	.135	287.760	1	.000	9.812	7.537	12.775
	Constant	-1.607	.092	305.300	1	.000	.201		

a. Variable(s) entered on step 1: sex.

Just for reference: $\ln(\frac{p}{1-p}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_j x_j + \beta_q x_q + \varepsilon_i$

30

Odds—How To

Y\X	female (1)	male (0)
Did not survive (0)	0.51	5.0
Survived (1)	1.97	0.20

Ln (y) = -1.607 + (sex * 2.284)

Ln (odds female survived) = -1.607 + (1 * 2.284) = .677
Ln (odds male survived) = -1.607

Odds female survived = Exp(.677) = 1.97
Odds male survived = Exp (-1.607) = 0.20

Odds female/Odds male = 1.97/0.20 = 9.85 -> Exp(b)

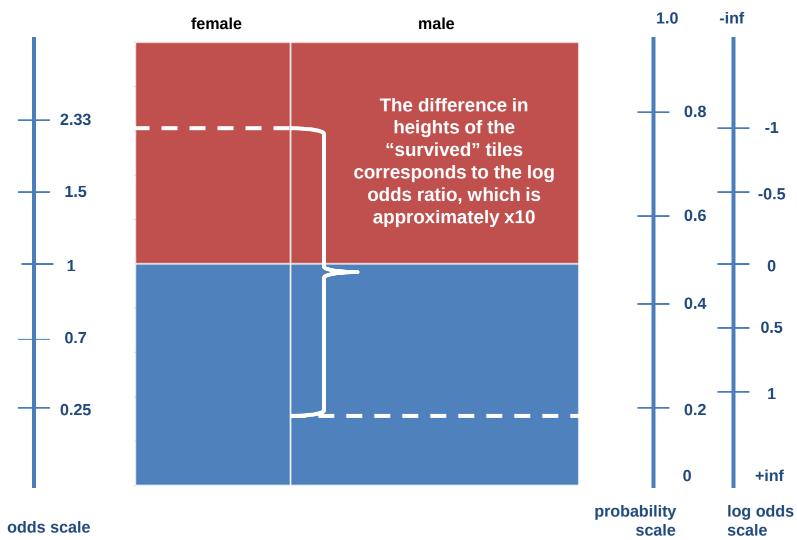
31

Probabilities, Odds, and Logits

P_i	$1 - P_i$	Odds $P_i/(1 - P_i)$	Logit	
.1	.9	.111	-2.20	**Stronger association**
.2	.8	.25	-1.39	
.3	.7	.429	-.847	
.4	.6	.667	-.405	**Weaker association**
.5	**.5**	**1**	**0**	**Independence**
.6	.4	1.5	.405	**Weaker association**
.7	.3	2.33	.847	
.8	.2	4	1.39	
.9	.1	9	2.20	**Stronger association**

32

Mosaic Plot with Odds



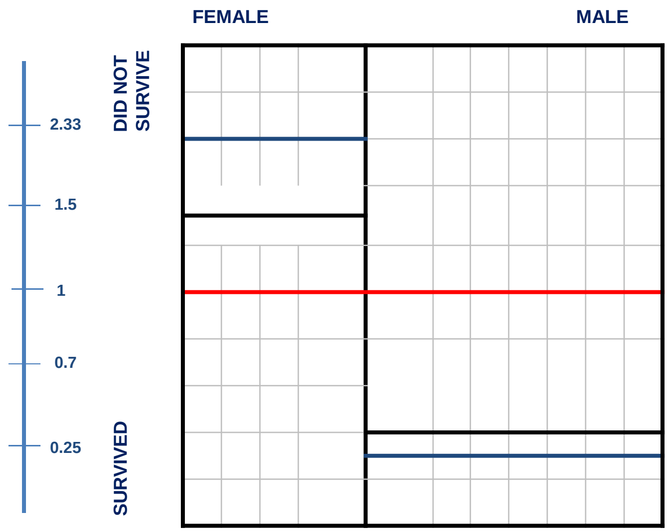
33

Plotting Odds Exercise

1. Draw a line of equal probability of survival
   – Odds = 1
2. Increase the height of surviving females tile
   – Odds = 2
3. Lower the height of the surviving males tile
   – Odds = 0.2
4. Check the difference in tile heights (~ x10)

peacecorps.gov

Plotting Odds—Check Your Worksheet



35

Major Criticisms

- Too much information in one chart
  - Simultaneous manipulation of heights and widths
- Log odds values beyond -2 or 2 can not be visually assessed

36

Questions

- Pratt Index
- Mosaic plots

37

Contact Information

Marina Murray

mmurray@peacecorps.gov

38