STAT 226 Lecture 14

Yibi Huang

Outline

- Models w/ Ordinal Explanatory Variables
- Models Allowing Interactions Btw Explanatory Variables

Horseshoe Crabs Data

```
crabs = read.table(
  "https://www.stat.uchicago.edu/~yibi/s226/horseshoecrabs.txt",
  header=TRUE
)
crabs$has.sate = as.numeric(crabs$Satellites>0)
```

Models w/ Ordinal Explanatory Variables

Models w/ Ordinal Explanatory Variables

- Recall `Color` of horseshoe crabs is ordinal (light to dark). Models with dummy variables treat color as nominal.
- To treat `Color` numerical, could assign *scores* such as (1,2,3,4) representing

  $1 =$ medium light, $2 =$ medium, $3 =$ medium dark, $4 =$ dark

  or the scores (1,1,2,4) representing

  $1 =$ medium light, $1 =$ medium, $2 =$ medium dark, $4 =$ dark

  or other scores, and then include the ***score* of Color as a numerical explanatory variable** in the model.

  $$\text{logit}(\pi) = \alpha + \gamma c + \beta x, \quad c\text{: color score, } x\text{: width}$$

Using the score *(1,2,3,4)*, controlling for width, odds of having satellite(s) become e^{γ} times as large for each 1-category increase in shell darkness.

Using the score *(1,2,3,4)*, controlling for width, odds of having satellite(s) become e^{γ} times as large for each 1-category increase in shell darkness.

Using the score *(1,1,2,4)*,

$$\text{odds} = \begin{cases} \exp(\alpha + \gamma + \beta x) & \text{if med. light or medium} \\ \exp(\alpha + 2\gamma + \beta x) & \text{if med. dark} \\ \exp(\alpha + 4\gamma + \beta x) & \text{if dark} \end{cases}$$

Controlling for width,

- no diff. in the odds of having satellite(s) between med.light- and medium crabs
- odds for med. dark crabs are e^{γ} times as high as for med.light and medium crabs
- odds for dark crabs are $e^{2\gamma}$ times as high as for med. dark

$$\text{med. light} \overset{same}{=} \text{medium} \xrightarrow{e^{\gamma}} \text{med. dark} \xrightarrow{e^{2\gamma}} \text{dark}$$

Ordinal Explanatory Variables, Different Scores

Same model as long as scores maintain the same relative spacings between categories

- so (1,2,3,4), (0,1,2,3), or (0,2,4,6) correspond to the same model
- but (1,2,3,5) is a different model

Using the scores (1,2,3,4):

```
crabs.fit3 = glm(has.sate ~ Color + Width, family=binomial, data=crabs)
summary(crabs.fit3)$coef
             Estimate Std. Error z value   Pr(>|z|)
(Intercept) -10.0708     2.8068  -3.588 0.00033326
Color        -0.5090     0.2237  -2.276 0.02286018
Width         0.4583     0.1040   4.406 0.00001053
```

Fitted model: $\text{logit}(\pi) = -10.071 - 0.509c + 0.458x$.

Controlling for width, odds of having satellite(s) is estimated to become $e^{\widehat{\gamma}} = e^{-0.509} = 0.601$ times as large for each 1-category increase in shell darkness.

Using the scores (1,1,2,4):

```
crabs$Cscore2 = crabs$Color
crabs$Cscore2[crabs$Color == 2] = 1
crabs$Cscore2[crabs$Color == 3] = 2
crabs.fit4 = glm(has.sate ~ Cscore2 + Width, family=binomial, data=crab
summary(crabs.fit4)$coef
             Estimate Std. Error z value    Pr(>|z|)
(Intercept) -10.6906     2.7360  -3.907 0.000093298
Cscore2      -0.4499     0.1760  -2.556 0.010580133
Width         0.4625     0.1045   4.427 0.000009546
```

- odds for medium dark crabs are estimated to be $\exp(\widehat{\gamma}) \approx \exp(-0.45) \approx 0.64$ times as high as for medium light and medium crabs are estimated to be
- odds for dark crabs are $\exp(2\gamma) \approx e^{2(-0.45)} = 0.41$ times as high compared to medium dark crabs

Does model treating color as nominal fit as well as model treating it as numerical with scores (1,2,3,4)?

H_0: $\text{logit}(\pi) = \alpha + \gamma c + \beta x$ (simpler (ordinal) model)

H_a: $\text{logit}(\pi) = \alpha + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x$ (more complex model)

```
crabs$C = as.factor(crabs$Color)
crabs.fit1 = glm(has.sate ~ C + Width, family=binomial, data=crabs)
anova(crabs.fit3, crabs.fit1, test="Chisq")
Analysis of Deviance Table

Model 1: has.sate ~ Color + Width
Model 2: has.sate ~ C + Width
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       170    189.121
2       168    187.457  2  1.66414  0.43515
```

LR stat = diff. in deviances = $189.12 - 187.46 = 1.66$

$df = 170 - 168 = 2$, P-value $= 0.4351$. Simpler model is adequate.

Does model treating color as nominal fit as well as model treating it as numerical with scores (1,1,2,4)?

```
anova(crabs.fit4, crabs.fit1, test="Chisq")
Analysis of Deviance Table

Model 1: has.sate ~ Cscore2 + Width
Model 2: has.sate ~ C + Width
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       170    187.658
2       168    187.457  2 0.200872  0.90444
```

LR stat = diff. in deviances = $187.66 - 187.46 = 0.2$

$df = 170 - 168 = 2$, P-value $= 0.9044$

Simpler model is adequate.

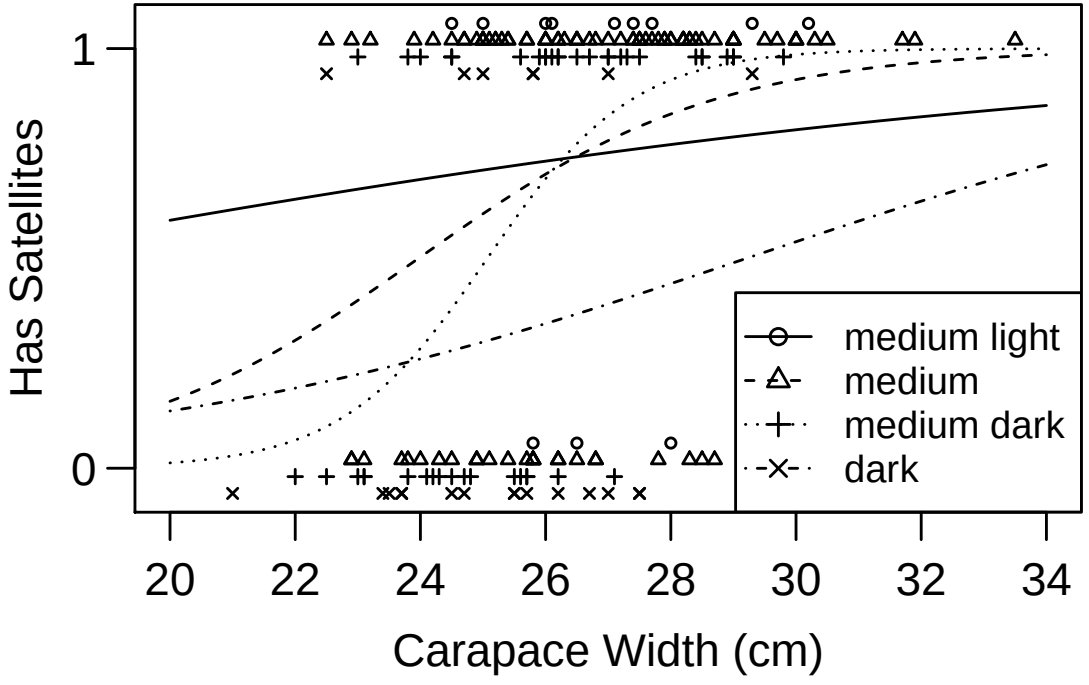
Models Allowing Interactions

Models Allowing Color*Width Interactions

- $\text{logit}(\pi) = \alpha + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x + \gamma_2 c_2 x + \gamma_3 c_3 x + \gamma_4 c_4 x$

$$= \begin{cases} \alpha + \beta x & \text{if medium light} \\ \alpha + \beta_2 + (\beta + \gamma_2)x & \text{if medium} \\ \alpha + \beta_3 + (\beta + \gamma_3)x & \text{if medium dark} \\ \alpha + \beta_4 + (\beta + \gamma_4)x & \text{if dark} \end{cases}$$

Different colors have different coefficient for "Width."



$$\text{odds} = \exp(\alpha + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x + \gamma_2 c_2 x + \gamma_3 c_3 x + \gamma_4 c_4 x)$$

$$= \begin{cases} \exp(\alpha + \beta x) & \text{if medium light} \\ \exp(\alpha + \beta_2 + (\beta + \gamma_2)x) & \text{if medium} \\ \exp(\alpha + \beta_3 + (\beta + \gamma_3)x) & \text{if medium dark} \\ \exp(\alpha + \beta_4 + (\beta + \gamma_4)x) & \text{if dark} \end{cases}$$

For every 1 cm increase in width, the odds of having satellite(s) become

- $\exp(\beta)$ times as large for medium light crabs
- $\exp(\beta + \gamma_2)$ times as large for medium crabs
- $\exp(\beta + \gamma_3)$ times as large for medium dark crabs
- $\exp(\beta + \gamma_4)$ times as large for dark crabs

$\Rightarrow$ `Width` effect changes with `Color`

— No homogeneous association

$$\text{odds} = \exp(\alpha + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x + \gamma_2 c_2 x + \gamma_3 c_3 x + \gamma_4 c_4 x)$$

$$= \begin{cases} \exp(\alpha + \beta x) & \text{if medium light} \\ \exp(\alpha + \beta_2 + (\beta + \gamma_2)x) & \text{if medium} \\ \exp(\alpha + \beta_3 + (\beta + \gamma_3)x) & \text{if medium dark} \\ \exp(\alpha + \beta_4 + (\beta + \gamma_4)x) & \text{if dark} \end{cases}$$

Controlling for `Width` $= x$,

$$\frac{\text{odds for medium crabs}}{\text{odds for med. light crabs}} = \frac{e^{\alpha+\beta_2+(\beta+\gamma_2)x}}{e^{\alpha+\beta x}} = \exp(\beta_2 + \gamma_2 x)$$

Similarly,

- odds for med. dark crabs are $\exp(\beta_3 + \gamma_3 x)$ times as large
- odds for dark crabs are $\exp(\beta_4 + \gamma_4 x)$ times as large

compared to med. light crabs.

$\Rightarrow$ `Color` effect changes with `Width` (x) — *No homo. association*

```
crabs.fit5 = glm(has.sate ~ C + Width + C*Width,
                 family=binomial, data=crabs)
summary(crabs.fit5)$coef
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.75261    11.4641 -0.1529   0.8785
C2           -8.28735    12.0036 -0.6904   0.4899
C3          -19.76545    13.3425 -1.4814   0.1385
C4           -4.10122    13.2753 -0.3089   0.7574
Width         0.10600     0.4266  0.2485   0.8037
C2:Width      0.31287     0.4479  0.6985   0.4849
C3:Width      0.75237     0.5043  1.4918   0.1358
C4:Width      0.09443     0.5004  0.1887   0.8503
```

Test of Interaction = Test of Homogeneous Association

Testing H_0: no interaction ($\gamma_2 = \gamma_3 = \gamma_4 = 0$)

```
anova(crabs.fit1,crabs.fit5,test="Chisq")
Analysis of Deviance Table

Model 1: has.sate ~ C + Width
Model 2: has.sate ~ C + Width + C * Width
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       168    187.457
2       165    183.081  3  4.37641  0.22358
```

LR stat = diff. in deviances = $187.46 - 183.08 = 4.3764$

$df = 168 - 165 = 3$, P-value $= 0.2236$

Simpler model is adequate (no interaction).

Models w/ Two Categorical Predictors & Their Interactions

Example: Smoking & Longivity Revisit

A survey during 1972-74 recruited 1314 women in the United Kingdom and asked if they smoked. Twenty years later, a follow-up survey determined whether each woman was deceased or still alive. The table below shows the result by the the women's age in the first survey (1972-74).

Age	18-34		35-54		55-64		65+	
	Dead	Alive	Dead	Alive	Dead	Alive	Dead	Alive
Smoker	5	174	41	198	51	64	42	7
Nonsmoker	6	213	19	180	40	81	165	28

Model: $\text{logit}(\pi) = \alpha + \beta x + \beta_{35} A_{35} + \beta_{55} A_{55} + \beta_{65} A_{65}$

$\pi = \text{P(Death)}$

$$x = \begin{cases} 1 & \text{if smoker} \\ 0 & \text{if nonsmoker} \end{cases}$$

$$A_{35} = \begin{cases} 1 & \text{if Age = 35-54} \\ 0 & \text{otherwise} \end{cases}$$

$$A_{55} = \begin{cases} 1 & \text{if Age = 55-64} \\ 0 & \text{otherwise} \end{cases}$$

$$A_{65} = \begin{cases} 1 & \text{if Age = 65+} \\ 0 & \text{otherwise} \end{cases}$$

Age	Smoker	logit(π)
18-34	N	α
	Y	$\alpha + \beta$
35-54	N	$\alpha + \quad + \beta_{35}$
	Y	$\alpha + \beta + \beta_{35}$
55-64	N	$\alpha + \quad + \beta_{55}$
	Y	$\alpha + \beta + \beta_{55}$
65+	N	$\alpha + \quad + \beta_{65}$
	Y	$\alpha + \beta + \beta_{65}$

Homogeneous Association

The model

$$\text{logit}(\pi) = \alpha + \beta x + \beta_{35} A_{35} + \beta_{55} A_{55} + \beta_{65} A_{65}$$

has **no interaction term**, which means the same conditional odds ratio

$$\frac{\text{odds for smokers}}{\text{odds for nonsmokers}} = \frac{e^{\alpha+\beta+\beta_{35}A_{35}+\beta_{55}A_{55}+\beta_{65}A_{65}}}{e^{\alpha+\beta_{35}A_{35}+\beta_{55}A_{55}+\beta_{65}A_{65}}} = e^{\beta}$$

for all 4 age groups. That is *homogeneous association* — same conditional odds ratio at each level of other variable.

Likewise, the conditional odds ratio for "Age" is also constant regardless of smoking status.

$$\frac{\text{odds for 35-54 age group}}{\text{odds for 18-34 age group}} = \frac{e^{\alpha+\beta x+\beta_{35}}}{e^{\alpha+\beta x}} = e^{\beta_{35}}$$

Age	18-34		35-54		55-64		65+	
	Dead	Alive	Dead	Alive	Dead	Alive	Dead	Alive
Smoker	5	174	41	198	51	64	42	7
Nonsmoker	6	213	19	180	40	81	165	28

```
Dead = c(5, 6, 41, 19, 51, 40, 42, 165)
Alive = c(174, 213, 198, 180, 64, 81, 7, 28)
Smoker = rep(c("Y","N"), 4)
Age = c("18-34","18-34","35-54","35-54","55-64","55-64","65+", "65+")
UKSmoke = data.frame(Smoker, Age, Dead, Alive)
UKSmoke
  Smoker   Age Dead Alive
1      Y 18-34    5   174
2      N 18-34    6   213
3      Y 35-54   41   198
4      N 35-54   19   180
5      Y 55-64   51    64
6      N 55-64   40    81
7      Y   65+   42     7
8      N   65+  165    28
```

```
fit1 = glm(cbind(Dead, Alive) ~ Smoker + Age,
           family = binomial, data=UKSmoke)
summary(fit1)$coef
            Estimate Std. Error z value  Pr(>|z|)
(Intercept)   -3.787     0.3212 -11.790 4.378e-32
SmokerY        0.450     0.1757   2.561 1.044e-02
Age35-54       1.683     0.3364   5.001 5.702e-07
Age55-64       3.096     0.3343   9.260 2.050e-20
Age65+         5.484     0.3635  15.088 1.945e-51
```

Controlling for Age, odds of death for smokers are estimated to be $e^{\widehat{\beta}} = e^{0.45} \approx 1.5684$ times the odds for nonsmokers.

95% Wald CI for e^{β}:

$$e^{\widehat{\beta} \pm 1.96 \times \text{SE}} \approx e^{0.45 \pm 1.96 \times 0.176} \approx (e^{0.106}, e^{0.794}) \approx (1.111, 2.213)$$

Significant adverse effect of smoking after accounting for Age.

95% Likelihood Ratio CIs for β & e^{β}:

```
confint(fit1, test="Chisq")
Waiting for profiling to be done...
                2.5 %  97.5 %
(Intercept) -4.4752 -3.2053
SmokerY      0.1087  0.7984
Age35-54     1.0625  2.3940
Age55-64     2.4821  3.8046
Age65+       4.8126  6.2466
exp(confint(fit1, test="Chisq"))
Waiting for profiling to be done...
                  2.5 %    97.5 %
(Intercept)   0.01139   0.04055
SmokerY       1.11487   2.22206
Age35-54      2.89368  10.95714
Age55-64     11.96609  44.90573
Age65+      123.04669 516.25855
```

At 95% confidence, the odds of death for smokers are 1.115 to 2.222 times the odds for nonsmokers in the same age group.

Estimation of Common Odds Ratio

- MH estimate of the common odds ratio (See Slides `L08.pdf`).
- In the logistic regression model:

$$\text{logit}(\pi) = \alpha + \beta x + \beta_{35} A_{35} + \beta_{55} A_{55} + \beta_{65} A_{65},$$

  e^{β} is the common odds ratio, and $e^{\widehat{\beta}}$ is the maximum likelihood estimate (MLE) for the common odds ratio. One can construct the Wald or LR confidence interval for e^{β} (See the previous two pages).
- MH estimate is preferred over MLE of the common odds ratio.

Tests of Conditional Independence

In the model

$$\text{logit}(\pi) = \alpha + \beta x + \beta_{35} A_{35} + \beta_{55} A_{55} + \beta_{65} A_{65},$$

$\beta = 0$ means conditional odds ratio $e^{\beta} = e^{0} = 1$, i.e., survival and smoking are **conditionally independent** given age.

Tests of conditional independence:

- CMH test
  - In fact, CMH test is the score test of $\beta = 0$ in the logistic model
- Wald test of $\beta = 0$ in the logistic model
- LR test of $\beta = 0$ in the logistic model

Wald test of conditional independence gives P-value ≈ 0.0104

```
summary(fit1)$coef
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)    -3.787     0.3212 -11.790 4.378e-32
SmokerY         0.450     0.1757   2.561 1.044e-02
Age35-54        1.683     0.3364   5.001 5.702e-07
Age55-64        3.096     0.3343   9.260 2.050e-20
Age65+          5.484     0.3635  15.088 1.945e-51
```

LR test of conditional independence gives P-value ≈ 0.0096:

```
drop1(fit1, "Smoker", test="Chisq")
Single term deletions

Model:
cbind(Dead, Alive) ~ Smoker + Age
       Df Deviance  AIC  LRT Pr(>Chi)
<none>        1.93 48.1
Smoker  1     8.64 52.8 6.71   0.0096
```

CMH test gives the P-value 0.0103 (See Week 4 Problem Session).

Comparison of the Three Tests of Conditional Independence

- The 3 tests usually agree when the sample sizes in each partial table are big enough
- Wald and LR tests require the sample size in each partial table to be large enough
- CMH test can work when the counts in the partial tables are small as long as the overall count is large enough
- In H_a, Wald and LR tests assume homogeneous association, but CMH test does not assume equality of odds ratios
- To sum up, for testing conditional independence in $2 \times 2 \times K$ tables, CMH test is preferred over Wald or LR tests.

Test of Homogeneous Association

The conditional odds ratios of smoking status and survival for the 4 age groups are as follows.

Age	18-34		35-54		55-64		65+	
	Dead	Alive	Dead	Alive	Dead	Alive	Dead	Alive
Smoker	5	174	41	198	51	64	42	7
Nonsmoker	6	213	19	180	40	81	165	28
Odds Ratio	$\frac{5\times213}{174\times6} \approx 1.02$		$\frac{41\times180}{198\times19} \approx 1.962$		$\frac{51\times81}{64\times40} \approx 1.614$		$\frac{42\times28}{7\times165} \approx 1.018$	

How to test if the 4 partial tables above have homogeneous association (identical conditional odds ratio)?

Test of Homogeneous Association

If we include the interaction term,

$$\text{Model 2: logit}(\pi) = \alpha + \beta x + \beta_{35}A_{35} + \beta_{55}A_{55} + \beta_{65}A_{65} + \gamma_{35}xA_{35} + \gamma_{55}xA_{65} + \gamma_{65}xA_{65},$$

the conditional odds ratio

$$\frac{\text{odds for Smokers}}{\text{odds for Nonsmokers}} = \frac{e^{\alpha+\beta+\beta_{35}A_{35}+\beta_{55}A_{55}+\beta_{65}A_{65}+\gamma_{35}xA_{35}+\gamma_{55}xA_{65}+\gamma_{65}xA_{65}}}{e^{\alpha+\beta_{35}A_{35}+\beta_{55}A_{55}+\beta_{65}A_{65}}}$$

$$= e^{\beta+\gamma_{35}xA_{35}+\gamma_{55}xA_{65}+\gamma_{65}xA_{65}}$$

changes with Age, if any of $\gamma_{35}, \gamma_{55}, \gamma_{65} \neq 0$.

H_0: $\gamma_{35} = \gamma_{55} = \gamma_{65} = 0$ means homogeneous association.

Test of Homogeneous Association

```
fit2 = glm(cbind(Dead, Alive) ~ Smoker + Age + Smoker*Age,
           family = binomial, data=UKSmoke)
anova(fit1, fit2, test="Chisq")
Analysis of Deviance Table

Model 1: cbind(Dead, Alive) ~ Smoker + Age
Model 2: cbind(Dead, Alive) ~ Smoker + Age + Smoker * Age
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         3     1.9264
2         0     0.0000  3   1.9264  0.58782
```

From the large P-value, we see no significant difference in the conditional odds ratios. The effect of smoking on the odds of death didn't change significantly with age.

Homogeneous Association v.s. Conditional Independence

To know whether Smoking and Survival were **homogeneously associated** given Age, i.e., whether the effect of Smoking on the odds of death changes with Age,

- test the significance of the interaction `Smoker*Age`.

To test whether Smoking and Survival were **conditionally independent** given Age, conduct a LRT test comparing the models

- `~ Smoker + Age`
- `~ Age`