STAT 226 Lecture 12-13

Section 4.4 Multiple Logistic Regression

Yibi Huang

Multiple Logistic Regression

Response: Y binary, $\quad \pi = \mathrm{P}(Y = 1)$

Explanatory variables: $x_1, x_2, \ldots, x_k$
can be numerical, categorical (dummy variables), or both.

Model form is

$$\text{logit}(\pi) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

or equivalently

$$\pi = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}$$

β_i = partial effect of x_i controlling for other variables in model

e^{β_i} = conditional odds ratio at $x_i + 1$ vs at x_i **keeping other x's fixed**

= multiplicative effect on odds of 1-unit increase in x_i

w/ other x's fixed

Adding a Categorical Explanatory Variable

Besides `Width` (X), add a categorical predictor — `Color`, coded as

$$1 = \text{medium light},\ 2 = \text{medium},\ 3 = \text{medium dark},\ 4 = \text{dark}$$

For a *categorical* predictor, need to create a **dummy variable** (= **indicator variable**) for each category:

$$c_1 = \begin{cases} 1 & \text{medium light} \\ 0 & \text{o/w} \end{cases}, \quad c_2 = \begin{cases} 1 & \text{medium} \\ 0 & \text{o/w} \end{cases},$$

$$c_3 = \begin{cases} 1 & \text{medium dark} \\ 0 & \text{o/w} \end{cases}, \quad c_4 = \begin{cases} 1 & \text{dark} \\ 0 & \text{o/w} \end{cases}$$

$$\boxed{\text{Model: } \text{logit}(\pi) = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x}$$

- $c_1 + c_2 + c_3 + c_4 = 1$ always true, so one of them is redundant.
- To account for redundancy, need to set one of $\beta_1, \beta_2, \beta_3, \beta_4$ to 0

Model 1:

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x$$

$$= \begin{cases} \alpha + \beta x & \text{if med. light } (c_2 = c_3 = c_4 = 0) \\ \alpha + \beta_2 + \beta x & \text{if medium } (c_2 = 1, c_3 = c_4 = 0) \\ \alpha + \beta_3 + \beta x & \text{if med. dark } (c_2 = 0, c_3 = 1, c_4 = 0) \\ \alpha + \beta_4 + \beta x & \text{if dark } (c_2 = c_3 = 0, c_4 = 1) \end{cases}$$

- Here we set $\beta_1 = 0$
- The category with no dummy var. in the model (or with coefficient $\beta_i = 0$) is called the baseline category. In Model 1, the baseline category is the color medium light (Color $= 1$).

Effect of Color Controlling for Width

Below "odds" = odds of having at least one satellite

$$\text{odds} = \frac{\pi}{1-\pi} = e^{\alpha+\beta_2 c_2+\beta_3 c_3+\beta_4 c_4+\beta x} = \begin{cases} e^{\alpha+\beta x} & \text{if med. light} \\ e^{\alpha+\beta_2+\beta x} & \text{if medium} \\ e^{\alpha+\beta_3+\beta x} & \text{if med. dark} \\ e^{\alpha+\beta_4+\beta x} & \text{if dark} \end{cases}$$

For female crabs of the same width,

$$\frac{\text{odds for a medium crab}(C=2)}{\text{odds for a medium light crab}(C=1)} = \frac{e^{\alpha+\beta_2+\beta x}}{e^{\alpha+\beta x}} = e^{\beta_2}$$

- Likewise,
  - e^{β_3} = odds ratio of (med. dark v.s. med. light)
  - e^{β_4} = odds ratio of (dark v.s. med. light)
- e^{β_i}'s are odds ratios of a category v.s. the baseline category (medium light), for crabs of the same width.

What about Medium v.s. Dark Crabs?

What about comparisons between non-baseline categories?
Like, medium (Color $= 2$) v.s. dark (Color $= 4$) crabs?

For medium and dark crabs of the same width, the odds ratio is

$$\frac{\text{odds for a medium crab}}{\text{odds for a dark crab}} = \frac{e^{\alpha+\beta_2+\beta x}}{e^{\alpha+\beta_4+\beta x}} = e^{\beta_2-\beta_4}.$$

Likewise

- $e^{\beta_4-\beta_3}$ = odds ratio of (dark v.s. med. dark)
- $e^{\beta_3-\beta_2}$ = odds ratio of (med. dark v.s. medium)

For all the above, we see the effect of `Color` does not change with `Width` (x) — *homogeneous association*

Effect of Width Controlling for Color

$$\text{Model 1:} \quad \text{odds} = \frac{\pi}{1-\pi} = e^{\alpha+\beta_2 c_2+\beta_3 c_3+\beta_4 c_4+\beta x}$$

For female crabs of same color but different width x_1, x_2,

$$\frac{\text{odds for crabs of Width } x_1}{\text{odds for crabs of Width } x_2} = \frac{e^{\alpha+\beta_2 c_2+\beta_3 c_3+\beta_4 c_4+\beta x_1}}{e^{\alpha+\beta_2 c_2+\beta_3 c_3+\beta_4 c_4+\beta x_2}} = e^{\beta(x_1-x_2)}$$

$\Rightarrow$ Controlling for `Color`, `Width` effect does NOT change with `Color`— *homogeneous association*.

As neither the effect of `Color` changes with `Width`,
nor the effect of `Width` change with `Color`,
we said Model 1 assumes *no interaction* betw. `Color` & `Width`.

First load the data and create the response

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

Then we fit the model with `Color` and `Width` as the predictors

```
crabs.fit0 = glm(has.sate ~ Color + Width, family=binomial, data=crabs)
crabs.fit0$coef
(Intercept)       Color       Width
   -10.0708     -0.5090      0.4583
```

- Something Wrong?

First load the data and create the response

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

Then we fit the model with `Color` and `Width` as the predictors

```
crabs.fit0 = glm(has.sate ~ Color + Width, family=binomial, data=crabs)
crabs.fit0$coef
(Intercept)       Color       Width
   -10.0708     -0.5090      0.4583
```

- Something Wrong?
- R regards `Color` as a numeric variable taking value 1-4, not categorical, no dummy variables are created

Numerical or Categorical?

Regarding `Color` as **numerical** taking values 1, 2, 3, and 4, the model becomes $\log(\text{odds}) = \alpha + \gamma\,\text{Color} + \beta x$, or

$$\text{odds} = \frac{\pi}{1-\pi} = e^{\alpha+\gamma\,\text{Color}+\beta x} = \begin{cases} e^{\alpha+\gamma+\beta x} & \text{if med. light (Color=1)} \\ e^{\alpha+2\gamma+\beta x} & \text{if medium (Color=2)} \\ e^{\alpha+3\gamma+\beta x} & \text{if med. dark (Color=3)} \\ e^{\alpha+4\gamma+\beta x} & \text{if dark (Color=4)} \end{cases}$$

The OR between adjacent categories of `Color`, controlling for `Width`, would be

	If Regarding Color as	
Odds ratio of	Numerical	Categorical
dark v.s. med. dark	e^{γ}	$e^{\beta_4-\beta_3}$
med. dark v.s. medium	e^{γ}	$e^{\beta_3-\beta_2}$
medium v.s. med. light	e^{γ}	e^{β_2}

as.factor()

- The command `as.factor()` tells R that `Color` is categorical and the dummy variables c_1, c_2, c_3, c_4 are created automatically
- By default, R drops the indicator c_1 for the lowest level

```
crabs$C = as.factor(crabs$Color)
crabs.fit1 = glm(has.sate ~ C + Width, family=binomial, data=crabs)
crabs.fit1$coef
(Intercept)          C2          C3          C4       Width
  -11.38519     0.07242    -0.22380    -1.32992     0.46796
```

The fitted model is

$$\text{logit}(\widehat{\pi}) = -11.39 + 0.07c_2 - 0.22c_3 - 1.33c_4 + 0.468x$$

```
crabs.fit1$coef
(Intercept)          C2          C3          C4       Width
  -11.38519     0.07242    -0.22380    -1.32992     0.46796
```

Fitted model:

$$\text{logit}(\widehat{\pi}) = -11.39 + 0.07c_2 - 0.22c_3 - 1.33c_4 + 0.468x$$

For a medium light female ($c_2 = c_3 = c_4 = 0$) of width $x = 25$ cm,

$$\widehat{\pi} = \frac{\exp(-11.39 + 0.468 \times 25)}{1 + \exp(-11.39 + 0.468 \times 25)} \approx 0.58$$

```
crabs.fit1$coef
(Intercept)          C2          C3          C4       Width
  -11.38519     0.07242    -0.22380    -1.32992     0.46796
```

Fitted model:

$$\text{logit}(\widehat{\pi}) = -11.39 + 0.07c_2 - 0.22c_3 - 1.33c_4 + 0.468x$$

For a medium light female ($c_2 = c_3 = c_4 = 0$) of width $x = 25$ cm,

$$\widehat{\pi} = \frac{\exp(-11.39 + 0.468 \times 25)}{1 + \exp(-11.39 + 0.468 \times 25)} \approx 0.58$$

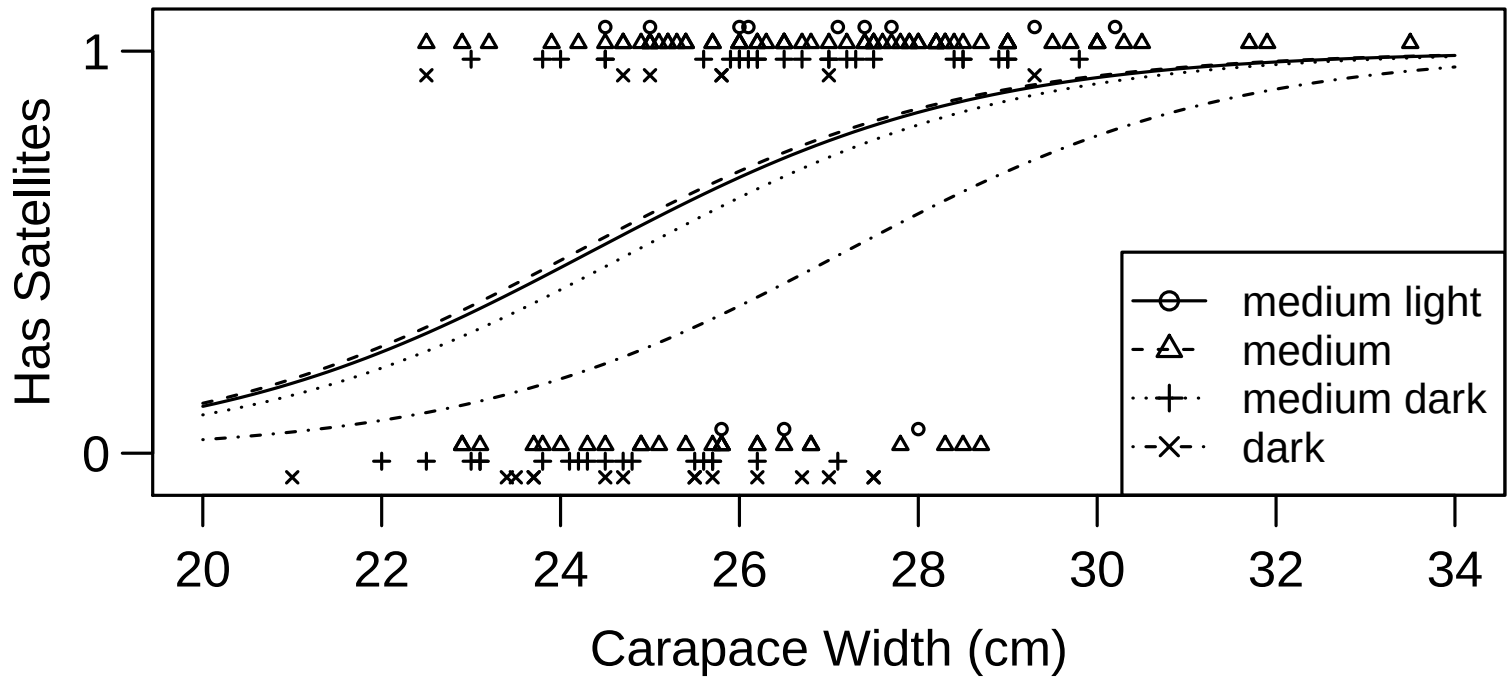
For a dark female ($c_2 = c_3 = 0, c_4 = 1$) of width $x = 25$ cm,

$$\widehat{\pi} = \frac{\exp(-11.39 + (-1.33)(1) + 0.468 \times 25)}{1 + \exp(-11.39 + (-1.33)(1) + 0.468 \times 25)} \approx 0.265.$$

$$\text{logit}(\widehat{\pi}) = -11.39 + 0.07c_2 - 0.22c_3 - 1.33c_4 + 0.468x$$

$$= \begin{cases} -11.39 + 0.468x & \text{if medium light} \\ -11.32 + 0.468x & \text{if medium} \\ -11.61 + 0.468x & \text{if medium dark} \\ -12.72 + 0.468x & \text{if dark} \end{cases}$$

Observe the four curves have the same shape because they have identical coefficient for Width.



Medium v.s. Medium Light Crabs (1)

```
summary(crabs.fit1)$coef
             Estimate Std. Error  z value    Pr(>|z|)
(Intercept) -11.38519     2.8735 -3.96219 0.000074264
C2            0.07242     0.7399  0.09787 0.922031566
C3           -0.22380     0.7771 -0.28800 0.773347793
C4           -1.32992     0.8525 -1.55998 0.118764113
Width         0.46796     0.1055  4.43373 0.000009262
```

Interpretation of $\widehat{\beta}_2 = 0.07242$: odds of having satellite(s) for medium crabs are estimated to be $e^{\widehat{\beta}_2} = e^{0.07242} \approx 1.07$ times the odds for medium light crabs of the same width.

Medium v.s. Medium Light Crabs

```
summary(crabs.fit1)$coef
             Estimate Std. Error  z value    Pr(>|z|)
(Intercept) -11.38519     2.8735 -3.96219 0.000074264
C2            0.07242     0.7399  0.09787 0.922031566
C3           -0.22380     0.7771 -0.28800 0.773347793
C4           -1.32992     0.8525 -1.55998 0.118764113
Width         0.46796     0.1055  4.43373 0.000009262
```

Under H_0: $\beta_2 = 0$, medium and medium light crabs do not differ in their chance of having satellite(s) given width. To test H_0: $\beta_2 = 0$, the Wald statistic is

$$z = \frac{\widehat{\beta_2}}{\text{SE}} = \frac{0.072}{0.74} = 0.098, \quad P\text{-value} = 0.922.$$

Conclusion: No significant diff. in the prob. of having satellites betw. Medium light and medium crabs of the same width.

Likelihood Ratio CI

95% LR CI for β_2 is $(-1.54, 1.45)$, which contains 0.

So LR test also fail to reject H_0: $\beta_2 = 0$.

```
confint(crabs.fit1,test="Chisq")
Waiting for profiling to be done...
                2.5 %  97.5 %
(Intercept) -17.3084 -5.9860
C2           -1.5397  1.4516
C3           -1.8919  1.2397
C4           -3.1357  0.2738
Width         0.2713  0.6870
```

What about (medium dark v.s. medium light) crabs?

What about (dark v.s. medium light) crabs?

What about Medium v.s. Dark Crabs?

For medium and dark crabs of the same width, the odds ratio is

$$\frac{\text{odds for a medium crab}}{\text{odds for a dark crab}} = \frac{e^{\alpha+\beta_2+\beta x}}{e^{\alpha+\beta_4+\beta x}} = e^{\beta_2-\beta_4}.$$

Estimated odds of having satellite(s) for a medium crab is

$$e^{\widehat{\beta}_2-\widehat{\beta}_4} = e^{0.07-(-1.33)} = e^{1.4} \approx 4.06$$

times the estimated odds for a dark crab of the same width.

What about Medium v.s. Dark Crabs?

For medium and dark crabs of the same width, the odds ratio is

$$\frac{\text{odds for a medium crab}}{\text{odds for a dark crab}} = \frac{e^{\alpha+\beta_2+\beta x}}{e^{\alpha+\beta_4+\beta x}} = e^{\beta_2-\beta_4}.$$

Estimated odds of having satellite(s) for a medium crab is

$$e^{\widehat{\beta}_2-\widehat{\beta}_4} = e^{0.07-(-1.33)} = e^{1.4} \approx 4.06$$

times the estimated odds for a dark crab of the same width.

However, to test $H_0 : \beta_2 = \beta_4$, need SE for $\widehat{\beta}_2 - \widehat{\beta}_4$, which is not provided in R.

What about Medium v.s. Dark Crabs?

For medium and dark crabs of the same width, the odds ratio is

$$\frac{\text{odds for a medium crab}}{\text{odds for a dark crab}} = \frac{e^{\alpha+\beta_2+\beta x}}{e^{\alpha+\beta_4+\beta x}} = e^{\beta_2-\beta_4}.$$

Estimated odds of having satellite(s) for a medium crab is

$$e^{\widehat{\beta}_2-\widehat{\beta}_4} = e^{0.07-(-1.33)} = e^{1.4} \approx 4.06$$

times the estimated odds for a dark crab of the same width.

However, to test $H_0 : \beta_2 = \beta_4$, need SE for $\widehat{\beta}_2 - \widehat{\beta}_4$, which is not provided in R.

The simplest solution is to change the baseline category. Say, use *dark* color as the baseline and model becomes

$$\text{Model 1a : } \text{logit}(\pi) = \alpha' + \beta_1' c_1 + \beta_2' c_2 + \beta_3' c_3 + \beta x$$

Change of Baseline

$$\text{Model 1} \;:\; \text{logit}(\pi) = \alpha \qquad + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x$$

$$\text{Model 1a} :\; \text{logit}(\pi) = \alpha' + \beta'_1 c_1 + \beta'_2 c_2 + \beta'_3 c_3 \qquad + \beta x$$

		logit(π) for	
Color	(c_1, c_2, c_3, c_4)	Model 1	Model 1a
med. light	$(1,0,0,0)$	$\alpha + \beta x$	$\alpha' + \beta'_1 + \beta x$
medium	$(0,1,0,0)$	$\alpha + \beta_2 + \beta x$	$\alpha' + \beta'_2 + \beta x$
med. dark	$(0,0,1,0)$	$\alpha + \beta_3 + \beta x$	$\alpha' + \beta'_3 + \beta x$
dark	$(0,0,0,1)$	$\alpha + \beta_4 + \beta x$	$\alpha' + \beta x$

The two models are equivalent, just a change of parameters.

$$\alpha' = \alpha + \beta_4, \quad \beta'_i = \beta_i - \beta_4 \quad \text{for } i = 1, 2, 3$$

Testing $\beta_2 = \beta_4$ in Model 1 is equivalent to testing $\beta'_2 = 0$ in Model 1a.

```
crabs$C1 = as.numeric(crabs$Color==1)
crabs$C2 = as.numeric(crabs$Color==2)
crabs$C3 = as.numeric(crabs$Color==3)
crabs.fit1a = glm(has.sate ~ C1+C2+C3 + Width, family=binomial,
                  data=crabs)
summary(crabs.fit1a)$coef
             Estimate Std. Error z value     Pr(>|z|)
(Intercept)  -12.715     2.7617  -4.604 0.000004144
C1             1.330     0.8525   1.560 0.118764113
C2             1.402     0.5484   2.557 0.010558984
C3             1.106     0.5921   1.868 0.061734755
Width          0.468     0.1055   4.434 0.000009262
```

- $\widehat{\beta}_2' = 1.4023$, which is equal to $\widehat{\beta}_2 - \widehat{\beta}_4$
- Wald test of H_0: $\beta_2' = 0$ gives P-value 0.0106

Conclusion: Medium and dark crabs of the same width differ significantly in the prob. of having satellites.

Likelihood Ratio Test (Medium v.s. Dark Crabs)

```
drop1(crabs.fit1a,test="Chisq")
Single term deletions

Model:
has.sate ~ C1 + C2 + C3 + Width
       Df Deviance AIC   LRT  Pr(>Chi)
<none>          188 198
C1      1       190 198  2.62    0.1058
C2      1       194 202  6.91    0.0086
C3      1       191 199  3.65    0.0560
Width   1       212 220 24.60 0.0000007
```

LR test of $\beta_2' = 0$ gives P-value 0.0086, same conclusion as Wald test.

Likelihood Ratio CI (Medium v.s. Dark Crabs)

```
confint(crabs.fit1a)
Waiting for profiling to be done...
                2.5 % 97.5 %
(Intercept) -18.45674 -7.579
C1           -0.27378  3.136
C2            0.35270  2.526
C3           -0.02792  2.314
Width         0.27128  0.687
```

95% for β_2' is $(0.353, 2.526)$

Interpretation: estimated odds for medium crabs are at least $e^{0.353} \approx 1.42$, at most $e^{2.526} \approx 12.5$ times the est. odds for dark crabs of the same width.

Likelihood Ratio Tests (LRT) for Model Comparison

Likelihood Ratio Tests can be used to compare between

- a simpler model, called the *reduced model*, and
- a more complex model, called the *full model*.

Likelihood Ratio Tests (LRT) for Model Comparison

Likelihood Ratio Tests can be used to compare between

- a simpler model, called the *reduced model*, and
- a more complex model, called the *full model*.

Note that

- the reduced model must be **a special case** of the full model.
  If not, CANNOT use LRT to do model comparison

Likelihood Ratio Tests (LRT) for Model Comparison

Likelihood Ratio Tests can be used to compare between

- a simpler model, called the *reduced model*, and
- a more complex model, called the *full model*.

Note that

- the reduced model must be **a special case** of the full model.
  If not, CANNOT use LRT to do model comparison
- H_0: the reduced model is correct
  H_a: the full model is correct, the reduced model is not

Likelihood Ratio Tests (LRT) for Model Comparison

Likelihood Ratio Tests can be used to compare between

- a simpler model, called the *reduced model*, and
- a more complex model, called the *full model*.

Note that

- the reduced model must be **a special case** of the full model. If not, CANNOT use LRT to do model comparison
- H_0: the reduced model is correct
  H_a: the full model is correct, the reduced model is not
- Rejecting H_0 means the reduced model doesn't fit the data well, compared to the full model

Likelihood Ratio Tests (LRT) for Model Comparison

Likelihood Ratio Tests can be used to compare between

- a simpler model, called the *reduced model*, and
- a more complex model, called the *full model*.

Note that

- the reduced model must be **a special case** of the full model. If not, CANNOT use LRT to do model comparison
- H_0: the reduced model is correct
  H_a: the full model is correct, the reduced model is not
- Rejecting H_0 means the reduced model doesn't fit the data well, compared to the full model
- Not rejecting H_0 means the reduced model fits the data nearly as well as the full model

Likelihood Ratio Test for Model Comparison

- Likelihood ratio (LR) statistic $= -2(L_0 - L_1)$, where
  $L_0 =$ max. log-likelihood for the reduced model,
  $L_1 =$ max. log-likelihood for the full model

Likelihood Ratio Test for Model Comparison

- Likelihood ratio (LR) statistic $= -2(L_0 - L_1)$, where
  $L_0 =$ max. log-likelihood for the reduced model,
  $L_1 =$ max. log-likelihood for the full model

- In general, $L_0 \leq L_1$.
- Under H_0, $L_0 \approx L_1$.

Likelihood Ratio Test for Model Comparison

- Likelihood ratio (LR) statistic $= -2(L_0 - L_1)$, where
  $L_0 =$ max. log-likelihood for the reduced model,
  $L_1 =$ max. log-likelihood for the full model

- In general, $L_0 \leq L_1$.
- Under H_0, $L_0 \approx L_1$.

- Large sample distribution of LR statistic is **Chi-squared** with

$$d.f. = \text{diff. in number of parameters for the 2 models}$$

Likelihood Ratio Test for Model Comparison

Rather than reporting the max. log-likelihood for a model, R reports

$$\text{Deviance} = -2(\text{max. log-likelihood} + C)$$

in which C is a constant depends only on the data but not the model. So

$$\begin{aligned}\text{LR statistic} &= -2(L_0 - L_1) \\ &= -2(L_0 + C) - [-2(L_1 + C)] \\ &= \text{diff. in deviance for the two models}\end{aligned}$$

- We will introduce deviance in Chapter 5
- d.f. for a deviance is

  (num. of observations) – (num. of parameters)
- so d.f. for a LR statistic = diff. in d.f. for the two deviances

```
> summary(crabs.fit1)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.38519    2.87346  -3.962 7.43e-05 ***
C2            0.07242    0.73989   0.098    0.922
C3           -0.22380    0.77708  -0.288    0.773
C4           -1.32992    0.85252  -1.560    0.119
Width         0.46796    0.10554   4.434 9.26e-06 ***
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.46  on 168  degrees of freedom
AIC: 197.46
```

For Model 1, deviance = "Residual deviance" = 187.46

d.f. of deviance = $173 - 5 = 168$

($n = 173$ for horseshoe crabs data)

Example: Likelihood Ratio Test of Color Effect Given Width

H_0 : $\text{logit}(\pi) = \alpha + \beta x$ (reduced model)

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

i.e., H_0 : $\beta_2 = \beta_3 = \beta_4 = 0$ (given width, Y indep. of color)

The `anova()` command in R can perform LRT comparing two models.

```
crabs.logit = glm(has.sate ~ Width, family = binomial, data=crabs)
anova(crabs.logit, crabs.fit1, test="Chisq")
Analysis of Deviance Table

Model 1: has.sate ~ Width
Model 2: has.sate ~ C + Width
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       171    194.453
2       168    187.457  3  6.99563 0.072037
```

Example: Likelihood Ratio Test of Color Effect Given Width (2)

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

Model 1: has.sate ~ Width
Model 2: has.sate ~ C + Width
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       171    194.453
2       168    187.457  3  6.99563 0.072037
```

LR statistic = diff. of deviance = $194.45 - 187.46 = 6.99$ with

$df = 171 - 168 = 3$, P-value= 0.072

$\Longrightarrow$ Some evidence (not strong) of Color effect given Width.

R command `drop1` on a model performs LRT comparing

H_0: the model w/ one term deleted v.s. H_a: the model itself

for each term in the model, e.g., P-value for `Width` in the R output below is LRT for comparing

$$H_0 : \text{logit}(\pi) = \alpha + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4$$
$$H_a : \text{logit}(\pi) = \alpha + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x$$

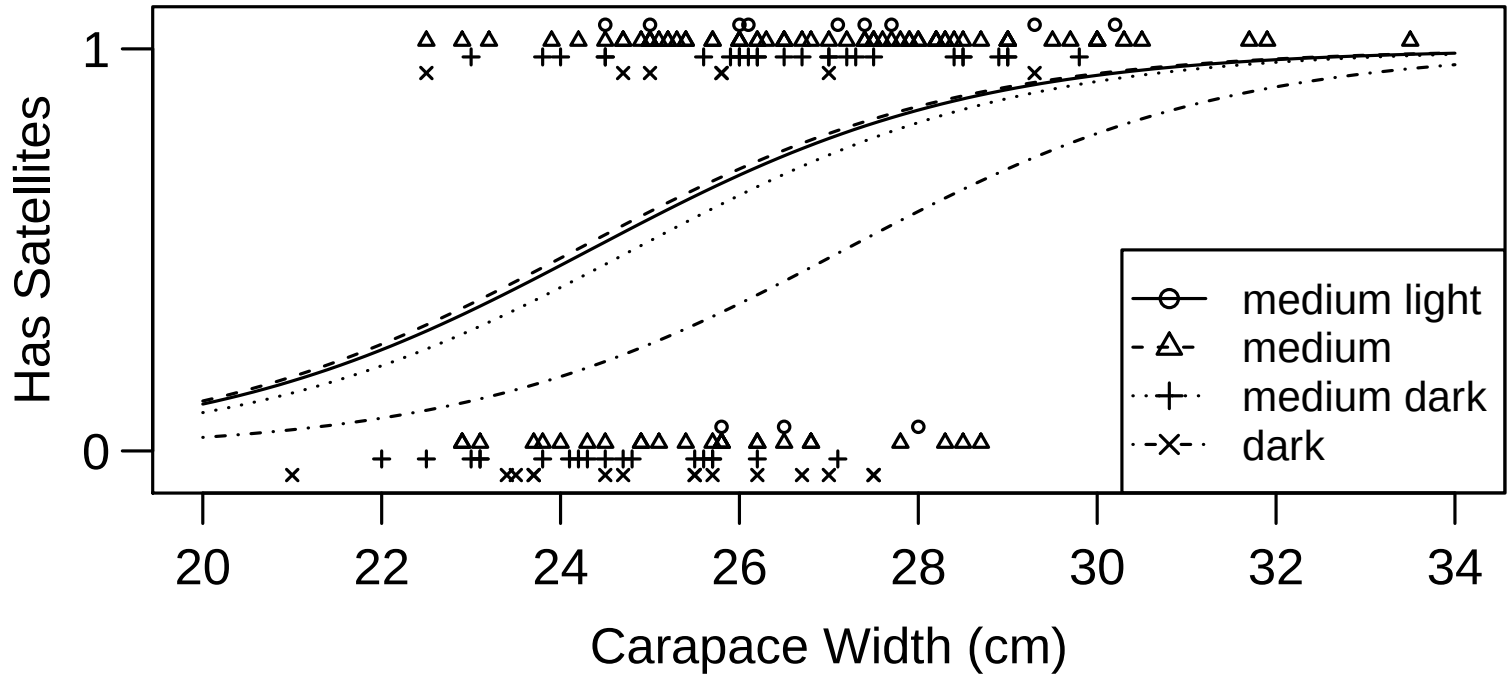
```
drop1(crabs.fit1, test="Chisq")
Single term deletions

Model:
has.sate ~ C + Width
       Df Deviance     AIC      LRT      Pr(>Chi)
<none>     187.457 197.457
C       3  194.453 198.453  6.99563      0.072037
Width   1  212.061 220.061 24.60381 0.00000070412
```

Some evidence (not strong) of `Color` effect given `Width`.

Strong evidence of `Width` effect given `Color`.

Other reduced models might be adequate.



From the plot of the four curves above, maybe only **dark** crabs are different from others.

Model 2: $\text{logit}(\pi) = \alpha + \beta_4 c_4 + \beta x$, where $c_4 = \begin{cases} 1 & \text{dark} \\ 0 & \text{o/w} \end{cases}$

```
crabs.fit2 = glm(has.sate ~ I(Color==4) + Width,
                 family=binomial, data=crabs)
summary(crabs.fit2)$coef
                      Estimate Std. Error    z value       Pr(>|z|)
(Intercept)       -11.67902562 2.69250507 -4.3376058 0.000014404319
I(Color == 4)TRUE  -1.30051207 0.52586104 -2.4731098 0.013394299762
Width               0.47822231 0.10414675  4.5918119 0.000004394143
```

Fitting gives $\widehat{\beta_4} = -1.300$ (SE $= 0.5259$).

Odds of satellites for a dark crab is estimated to be $e^{-1.300} = 0.27$ times the odds for a non-dark crab of the same width.

Compare model with 1 dummy for color to full model with 3 dummies.

H_0: $\text{logit}(\pi) = \alpha + \beta_4 c_4 + \beta x$ (reduced model)

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

Note H_0 is $\beta_2 = \beta_3 = 0$ in full model.

```
anova(crabs.fit2, crabs.fit1, test="Chisq")
Analysis of Deviance Table

Model 1: has.sate ~ I(Color == 4) + Width
Model 2: has.sate ~ C + Width
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       170    187.958
2       168    187.457  2 0.500847  0.77847
```

LR stat = diff. in deviances = $187.96 - 187.45 = 0.50$

$df = 170 - 168 = 2$, P-value $= 0.7785$. $\Rightarrow$ reduced model is adequate.

Ordinal Factors

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

Model 3: $\text{logit}(\pi) = \alpha + \gamma c + \beta x$, c: color, x : width

```
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 decrease by a factor of $e^{\widehat{\gamma}} = e^{-0.509} = 0.601$ for each 1-category increase in shell darkness.

Does model treating color as nominal fit as well as model treating it as qualitative?

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

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

```
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      189.1
2       168      187.5  2    1.664    0.435
```

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

$df = 170 - 168 = 2$, P-value $= 0.4351$

reduced model is adequate.