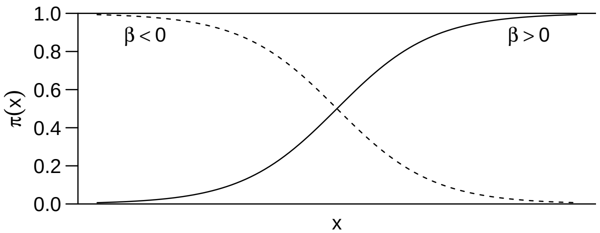
Section 4.1-4.2 Simple Logistic Regression

Simple logistic regression has a single explanatory variable x and models the success probability $\pi(x)$ for the binomial response as

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}.$$



- If $\beta = 0$, then $\pi(x) = \frac{e^{\alpha}}{1+e^{\alpha}}$ doesn't change with x
- bigger $|\beta|$, steeper curve
- **point of symmetry**:

$$\begin{aligned}\pi(x) = 1/2 &\iff e^{\alpha+\beta x} = 1 = e^0 \\ &\iff \alpha+\beta x = 0 \iff x = -\alpha/\beta.\end{aligned}$$

4.1.1 Linear Approximation Interpretations

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}, \quad \Rightarrow \quad 1-\pi(x) = \frac{1}{1+e^{\alpha+\beta x}}$$

One can show that

$$\frac{d}{dx}\pi(x) = \frac{\beta e^{\alpha+\beta x}}{(1+e^{\alpha+\beta x})^2} = \beta\pi(x)(1-\pi(x)).$$

i.e., **the slope of $\pi(x)$ at x is** $\boxed{\beta\pi(x)(1-\pi(x))}$.

- At x with $\pi(x) = \frac{1}{2}$, slope $= \beta \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{\beta}{4}$.
- At x with $\pi(x) = 0.1$ or 0.9, slope $= \beta \cdot 0.1 \cdot 0.9 = 0.09\beta$.
- **Steepest slope** at where $\pi(x) = 1/2$, i.e., **at point of symmetry** $x = -\alpha/\beta$.
- If x increases by Δx, then π increases by $\approx \beta\pi(1-\pi)\Delta x$.

Example: Horseshoe Crabs

See Section 3.3.2 and 4.1.2-4.1.3 for background information.

```
> crabs = read.table("horseshoecrabs.dat", header = T)
> crabs
    Color Spine Width Weight Satellites
1       2     3  28.3  3.050          8
2       3     3  22.5  1.550          0
3       1     1  26.0  2.300          9
4       3     3  24.8  2.100          0
5       3     3  26.0  2.600          4
6       2     3  23.8  2.100          0
... (omitted) ...
173     2     2  24.5  2.000          0
```


Example: Horseshoe Crabs

$$Y = \begin{cases} 1 & \text{if female crab has satellite(s)} \\ 0 & \text{if no satellites} \end{cases}$$

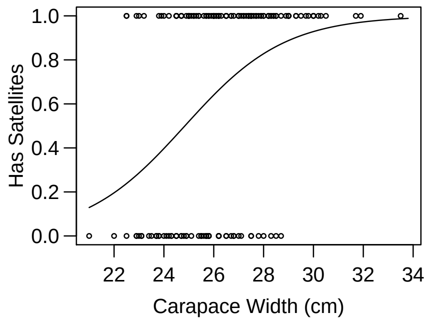
$$X = \text{carapace width (cm) of female crab}$$

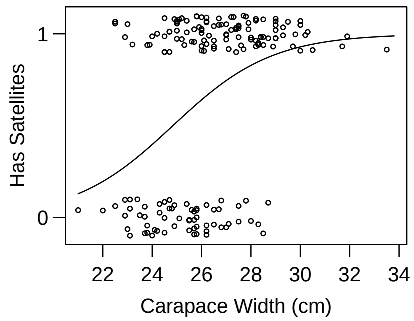
```
> attach(crabs)
> has.sate = as.numeric(Satellites > 0)
> crabs.logit = glm(has.sate ~ Width, family = binomial)
> crabs.logit$coef
(Intercept)       Width
-12.3508177   0.4972306
```

If unspecified, R use *logit* link by default. The fitted model is

$$\widehat{\pi}(x) = \frac{e^{-12.351+0.497x}}{1+e^{-12.351+0.497x}}$$

```
> plot(Width, has.sate,
     xlab = "Carapace Width (cm)",ylab = "Has Satellites")
> curve(exp(-12.351+0.497*x)/(1+exp(-12.351+0.497*x)), add = T)
```





There are multiple observations (crabs) at same points (left plot).

To see them, we can "jitter" their Y values by adding a small amount of noise (right plot).

```
> plot(Width, jitter(has.sate),
     xlab = "Carapace Width (cm)",ylab = "Has Satellites")
> curve(exp(-12.351+0.497*x)/(1+exp(-12.351+0.497*x)), add = T)
```

Hard to visually assess how well the curve fits the data

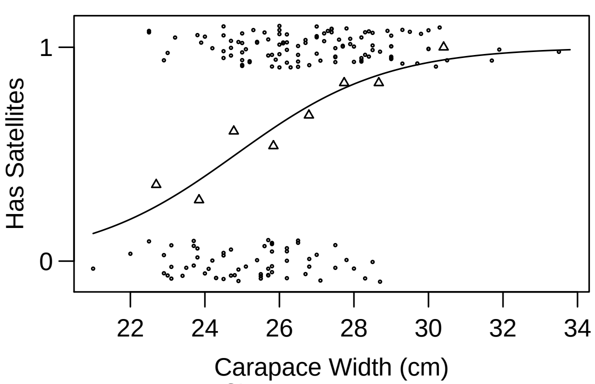
To better access the fit visually, one can group crabs of similer width and compute sample proportions for each group.

```
> wd.grp = cut(Width, breaks= c(0,23.25,24.25,25.25,26.25,
                               27.25,28.25,29.25,Inf))
> wd.table = table(wd.grp, Satellites > 0)
> wd.table

wd.grp        FALSE TRUE
  (0,23.2]        9    5
  (23.2,24.2]    10    4
  (24.2,25.2]    11   17
  (25.2,26.2]    18   21
  (26.2,27.2]     7   15
  (27.2,28.2]     4   20
  (28.2,29.2]     3   15
  (29.2,Inf]      0   14
> percent = wd.table[,2]/rowSums(wd.table)
> percent
   (0,23.2] (23.2,24.2] (24.2,25.2] (25.2,26.2] (26.2,27.2]
  0.3571429   0.2857143   0.6071429   0.5384615   0.6818182
(27.2,28.2] (28.2,29.2]  (29.2,Inf]
  0.8333333   0.8333333   1.0000000
```


The 8 triangle dots indicate the sample proportions against the mean widths of crabs in the 8 categories.

```
> wd.ave = tapply(Width,wd.grp,mean)
> wd.ave
   (0,23.2] (23.2,24.2] (24.2,25.2] (25.2,26.2] (26.2,27.2]
   22.69286    23.84286    24.77500    25.83846    26.79091
(27.2,28.2] (28.2,29.2]  (29.2,Inf]
   27.73750    28.66667    30.40714
> plot(Width, jitter(has.sate), cex=0.5,
     xlab = "Carapace Width (cm)",ylab = "Has Satellites")
> curve(exp(-12.351+0.497*x)/(1+exp(-12.351+0.497*x)), add = T)
> points(wd.ave, percent, pch = 2)   # "pch=2" use triangle dots
```



Fitted Model:

$$\widehat{\pi}(x) = \frac{\exp(\widehat{\alpha} + \widehat{\beta}x)}{1 + \exp(\widehat{\alpha} + \widehat{\beta}x)} = \frac{\exp(-12.351 + 0.497x)}{1 + \exp(-12.351 + 0.497x)}$$

- $\widehat{\beta} = 0.497 > 0$, so $\widehat{\pi}$ increases as Width (x) increases
- Point of symmetry:

$$\widehat{\pi}(x) = \frac{1}{2} \text{ when } x = -\frac{\widehat{\alpha}}{\widehat{\beta}} = -\frac{-12.351}{0.497} = 24.85\,\text{cm}$$

- Steepest slope at point of symmetry $x = 24.85$ cm with slope

$$\widehat{\beta}\pi(1-\pi) = 0.497 \times \frac{1}{2} \times \frac{1}{2} \approx 0.124$$

  If Width (x) increases by 1 cm, then π increases by 0.124 (actual $\widehat{\pi}$ at $x = 25.85$ is 0.623).
- At $x = 33.5$ (max. obs. width), $\widehat{\pi} \approx 0.987$, and the estimated slope is $0.497 \cdot (0.987) \cdot (1 - 0.987) \approx 0.0064$.

  $\Rightarrow$ Rate of change varies with x.

Predictions

The probability that an average-size female crab (w/ Width at $\overline{x} = 26.3$ cm) has satellite(s) is estimated to be

$$\widehat{\pi}(x) = \frac{e^{-12.351+0.497\times 26.3}}{1+e^{-12.351+0.497\times 26.3}} \approx 0.67$$

R provides two kinds of predicted values.

The first one gives $\widehat{\alpha} + \widehat{\beta}x = -12.351 + 0.497 \times 26.3 \approx 0.72.$

```
> predict(crabs.logit, data.frame(Width=26.3),type="link")
        1
0.7263467
```

The second one gives $\widehat{\pi}(x) = \frac{\exp(\widehat{\alpha}+\widehat{\beta}x)}{1+\exp(\widehat{\alpha}+\widehat{\beta}x)}$ as computed above.

```
> predict(crabs.logit, data.frame(Width=26.3),type="response")
        1
0.6740031
```


Remarks

- Fitting linear probability model $\pi(x) = \alpha + \beta x$ (binomial w/ identity link) fails in the crabs example.

```
> glm(has.sate ~ Width, family=binomial(link="identity"))
Error: no valid set of coefficients has been found:
please supply starting values
```

- If we pretend $Y \sim$ Normal and fit a linear regression model

$$Y = \alpha + \beta x + \varepsilon,$$

```
> lm(has.sate ~ Width)
Coefficients:
(Intercept)        Width
   -1.76553      0.09153
```

We get the model $\widehat{Y} = -1.7655 + 0.09153x$.
At $x = 33.5$ cm, the predicted value (estimated prob. of satellites) is

$$-1.7655 + 0.09153 \times 33.5 = 1.30 \quad !?!$$

Odds Ratio Interpretation of Logistic Models

Since $\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x$, odds are

$$\frac{\pi}{1-\pi} = \begin{cases} e^{\alpha+\beta x} & \text{at } x \\ e^{\alpha+\beta(x+1)} = e^{\beta}e^{\alpha+\beta x} & \text{at } x+1 \end{cases}$$

So

$$\frac{\text{odds at } (x+1)}{\text{odds at } x} = \frac{e^{\beta}\cancel{e^{\alpha+\beta x}}}{\cancel{e^{\alpha+\beta x}}} = e^{\beta}$$

More generally,

$$\frac{\text{odds at } (x+\Delta x)}{\text{odds at } x} = \frac{e^{\Delta x}\cancel{e^{\alpha+\beta x}}}{\cancel{e^{\alpha+\beta x}}} = e^{\beta\Delta x}$$

If $\beta = 0$, then $e^{\beta} = 1$ and odds do not depend on x.

Example (Horseshoe Crabs)

$$\widehat{\beta} = 0.497 \quad \Longrightarrow \quad e^{\widehat{\beta}} = e^{0.497} \approx 1.64.$$

Odds of having satellite(s) are estimated to increase by a factor of 1.64 for each 1 cm increase in width.

If width increases by 0.1 cm, then odds are estimated to increase by a factor of

$$e^{(0.497)(0.1)} = e^{0.0497} = 1.051.$$

Inference for Simple Logistic Regression

Wald CIs

Wald $(1-\alpha)100\%$ CIs for β are $\widehat{\beta} \pm z_{\alpha/2}\text{SE}(\widehat{\beta})$.

Example (Horseshoe Crabs)

```
> summary(crabs.logit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3508     2.6287  -4.698 2.62e-06 ***
Width         0.4972     0.1017   4.887 1.02e-06 ***
```

95% CI for β:

$$0.497 \pm (1.96)(0.102) = 0.497 \pm 0.200 = (0.297, 0.697)$$

95% CI for e^{β}: $(e^{0.297}, e^{0.697}) = (1.35, 2.01)$

$\Longrightarrow$ Odds are estimated to increase by a factor of at least 1.35, at most 2.01 for every 1 cm increment in Width.

Wald CI for β and e^{β} in R:

```
> confint.default(crabs.logit)
                  2.5 %     97.5 %
(Intercept) -17.5030100 -7.1986254
Width         0.2978326  0.6966286
> exp(confint.default(crabs.logit))
                   2.5 %       97.5 %
(Intercept) 2.503452e-08 0.0007476128
Width       1.346936e+00 2.0069749360
```

Safer to use **LR CI** than Wald CI.
For crabs example, 95% LR CI for e^{β} is $(1.36, 2.03)$.

```
> confint(crabs.logit)
Waiting for profiling to be done...
                  2.5 %     97.5 %
(Intercept) -17.8100090 -7.4572470
Width         0.3083806  0.7090167
> exp(confint(crabs.logit))
Waiting for profiling to be done...
                   2.5 %       97.5 %
(Intercept) 1.841668e-08 0.0005772432
Width       1.361219e+00 2.0319922986
```


Wald Tests for β

H_0: $\beta = 0$ (i.e., Y indep. of X, i.e., $\pi(x)$ constant in x)
H_a: $\beta \neq 0$

```
> summary(crabs.logit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3508     2.6287  -4.698 2.62e-06 ***
Width         0.4972     0.1017   4.887 1.02e-06 ***
```

We see

$$z = \frac{\widehat{\beta}}{\text{SE}} = \frac{0.4972}{0.1017} = 4.887$$

or

$$z^2 = 4.887^2 \approx 23.88, \quad df = 1 \quad \text{(chi-squared)}$$

P-value < 0.0001: very strong evidence that $\pi \uparrow$ with width.

Likelihood Ratio Tests for β

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

Model:
has.sate ~ Width
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>      194.45 198.45
Width   1   225.76 227.76 31.306 2.204e-08 ***
```