Categorical Data Analysis
Lecture 7 & 8 & 9

Graduate School of Advanced Science and Engineering
Rei Monden

2024.12.24

Logistic regression

Logistic regression

As discussed before, logistic regression is used when the outcome variable is categorical, binary (*success*, *failure*).

Today we will learn more about this particular GLM.

Logistic regression

The logistic regression model is a GLM, with:

- Response variable Y: Binary.
- *Random component:* Binomial distribution.
- *Link function:* Logit.

Logistic regression – One predictor

$$\underbrace{\log\left(\frac{\pi}{1-\pi}\right)}_{\text{logit}(\pi)} = \alpha + \beta x$$

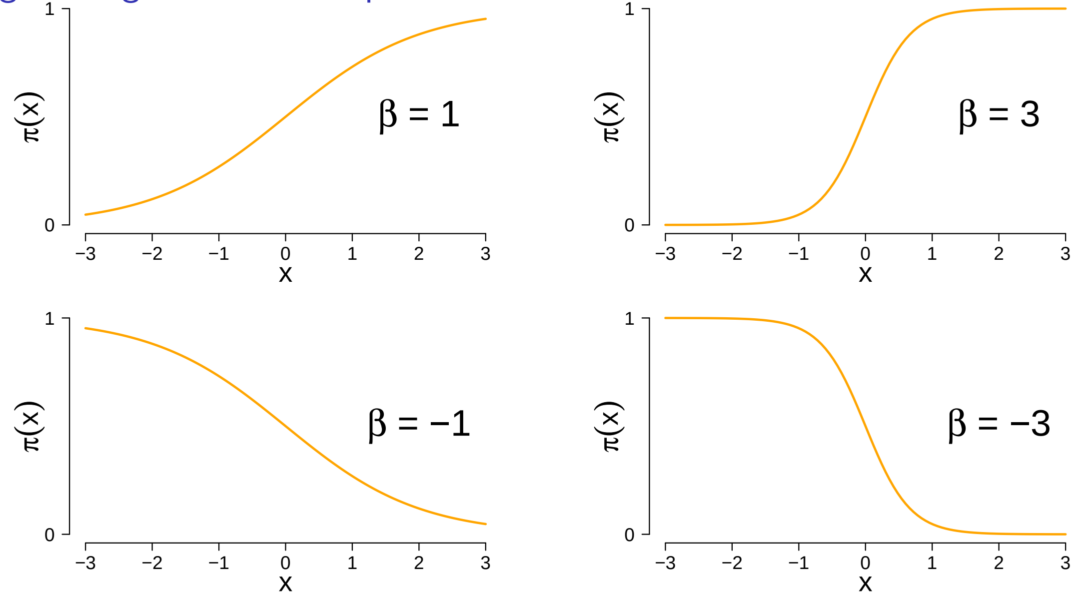
$\pi = P(Y = 1)$ is the probability of a success.
Note that π is a function of x: $\pi = \pi(x)$.
This can be seen by writing the model in terms of π:

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

$\pi(x)$ is an S-shaped curve, which:

- Increases (decreases) when $\beta > 0$ ($\beta < 0$).
- Is steeper as $|\beta|$ increases.

Logistic regression – One predictor



Logistic regression – Three interpretations

We will see three alternative ways of interpreting β, the regression effect of x on $\pi(x)$.

Additive interpretation (log odds)

$$\boxed{\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x}.$$

The log odds of a success are added by β per 1 unit increase of x.

Not very useful since we are not accustomed to "log odds".

Logistic regression – Three interpretations

Multiplicative interpretation (odds)

Exponentiating both sides of $\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x$ we have that

$$\boxed{\frac{\pi}{1-\pi} = \exp\left(\alpha + \beta x\right) = e^{\alpha}(e^{\beta})^{x}}$$

The odds of a success are multiplied by e^{β} per 1 unit increase of x.

This means that

$$\underbrace{e^{\beta}}_{\text{odds ratio}} = \frac{\text{odds at } (x+1)}{\text{odds at } x}.$$

Logistic regression – Three interpretations

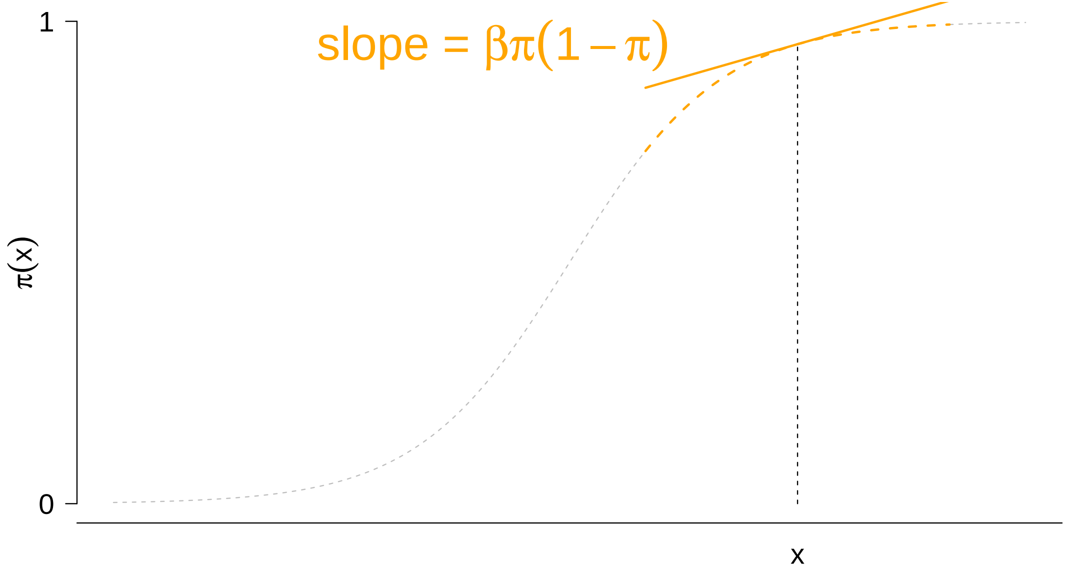
$\pi(x)$ itself

By means of approximating the S-shaped curve of $\pi(x)$ at some value x by the tangent at x (i.e., first derivative line), we use this derivative (rate of change) in the neighborhood of x.

$$\pi'(x) = \beta\pi(x)[1 - \pi(x)].$$

In the neighborhood of x, $\pi(x)$ is added by $\beta\pi(x)[1 - \pi(x)]$ per 1 unit increase of x.

Logistic regression – Three interpretations



Logistic regression – Example in R

width	y
28.3	1
22.5	0
26.0	1
⋮	⋮
28.0	0
27.0	0
24.5	0

173 rows in total.

Predictor: *width*, shell width in cm (continuous variable).
Outcome: *y*, binary (0 = female has no satellites; 1 = has satellites).

```r
# Import data frame from file:
crab.df <- read.table("Crabs.dat", header = TRUE)
```

Logistic regression – Example in R

Fit logistic regression:

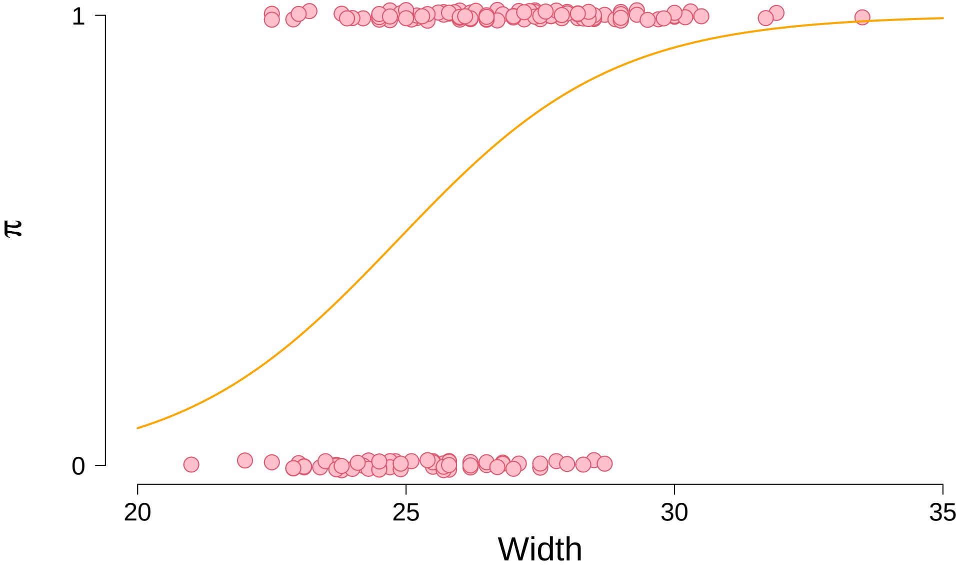
```
crab.fit <- glm(y ~ width,
                family = binomial(link = "logit"),
                data   = crab.df
                )

summary(crab.fit)
```

Output:

```
---------------------------
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 ***
---------------------------
```

Logistic regression – Example in R



Logistic regression – Example in R

$$\log\left(\frac{\pi}{1-\pi}\right) = -12.35 + 0.50x.$$

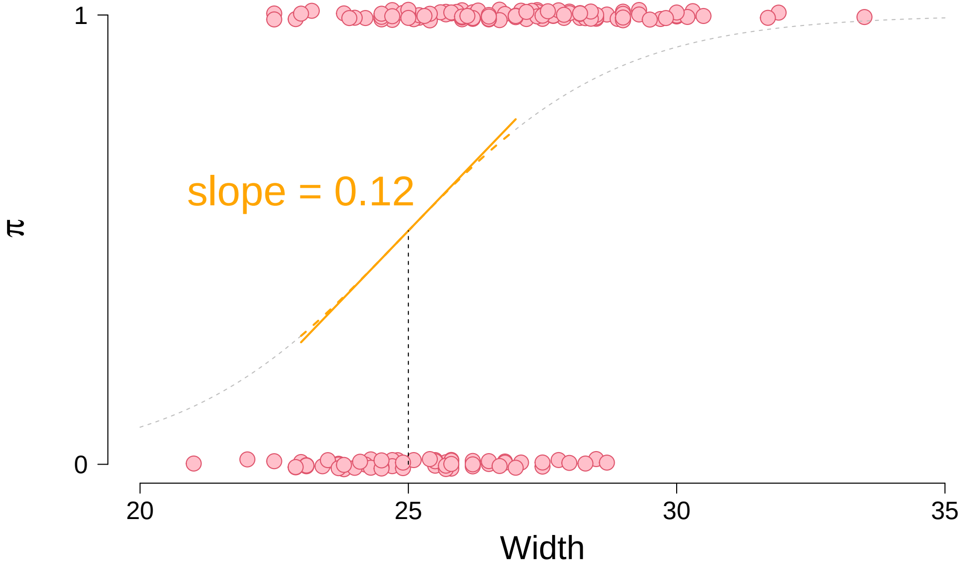
Interpreting the effect of x on $\pi = P(Y = 1)$:

- *Additive:*
  The estimated log odds of having at least one satellite increase 0.50 per 1cm increase of shell width.
- *Multiplicative:*
  The estimated odds of having at least one satellite are multiplied by $e^{0.50} = 1.64$ per 1cm increase of shell width.
- $\pi(x)$ *by linear approximation:*
  For example, noting that $\hat{\pi}(25) = .52$, we can say that, at width $= 25$, the estimated probability of having at least one satellite increases at the rate of

$$\beta\pi(x)\left[1-\pi(x)\right] = 0.50(.52)(1-.52) = 0.12$$

per 1cm increase of shell width.

Logistic regression – Example in R



Exercise 7-1

A study investigated characteristics associated with whether a cancer patient achieved remission (variable y, scored 1 = yes, 0 = no). An important explanatory variable was a labeling index (*LI* = percentage of "labeled" cells) that measures proliferative activity of cells after a patient receives an injection of tritiated thymidine. Below is the code to analyze the obtained data. Run the code below in R and answer the following questions.

```
LI         <- c(8, 8, 10, 10, rep(c(12, 14, 16), each = 3),
                18, 20, 20, 20, 22, 22, 24, 26, 28, 32, 34, rep(38, 3))
y          <- c(rep(0, 13), 1, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0)
cancer.fit <- glm(y ~ LI, family = binomial)
summary(cancer.fit)
confint(cancer.fit)
```

a. Show that $\hat{P}(Y = 1) = 0.50$ when $LI = 26.0$.
b. When LI increases by 1, show that the estimated odds of remission multiply by 1.16.
c. Describe the effect of LI on the estimated log-odds of remission.
d. Show that the rate of change in $\hat{P}(Y = 1)$ is 0.026 when $LI = 18$.

Statistical inference for logistic regression

Statistical inference for logistic regression

We will mostly use the Wald and likelihood-ratio methods that we learned before.

Wald test ($\mathcal{H}_0 : \beta = 0$) and CI for β

$$z = \frac{\hat{\beta}}{SE} \underset{\mathcal{H}_0}{\sim} \mathcal{N}(0,1), \qquad \hat{\beta} \pm z_{\alpha/2}(SE).$$

Wald CI for the odds ratio e^{β}

$$\exp\left(\hat{\beta} \pm z_{\alpha/2}(SE)\right).$$

As discussed in an earlier lecture, it is best to rely on the profile likelihood CI (i.e., likelihood-ratio test) when:

- The sample size n is small.
- $\hat{\pi}$ is close to 0 or 1.

Statistical inference for logistic regression

```
---------------------------
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 ***
---------------------------
```

Wald test for $\mathcal{H}_0 : \beta = 0$:

$$z_W = \frac{\hat{\beta}_{\text{width}}}{SE(\hat{\beta}_{\text{width}})} = 4.9.$$

Since $p < .05$, at significance level $\alpha = .05$, we reject the null hypothesis that $\beta_{\text{width}} = 0$.

Statistical inference for logistic regression

Wald 95% CI for β:

```
confint.default(crab.fit)
```

Output:

```
                  2.5 %    97.5 %
width         0.2978326 0.6966286
```

Wald 95% CI for e^{β}:

```
exp( confint.default(crab.fit) )
```

Output:

```
                  2.5 %    97.5 %
width       1.346936e+00 2.0069749360
```

Statistical inference for logistic regression

Likelihood-ratio test for $\mathcal{H}_0 : \beta = 0$:

```
drop1(crab.fit, test = "LRT")
```

Output:

```
----------------------------------------------
       Df Deviance    AIC    LRT  Pr(>Chi)    
width   1   225.76 227.76 31.306 2.204e-08 ***
----------------------------------------------
```

$\chi^2(1) = 31.306$, $p < .001$:
At significance level $\alpha = .05$, we reject the null hypothesis that $\beta_{\text{width}} = 0$.

Statistical inference for logistic regression

Likelihood-ratio CI, i.e., profile likelihood CI, for β:

```
confint(crab.fit)
```

Output:

```
                2.5 %    97.5 %
width       0.3083806 0.7090167
```

Likelihood-ratio CI, i.e., profile likelihood CI, for e^{β}:

```
exp( confint(crab.fit) )
```

Output:

```
                2.5 %      97.5 %
width    1.361219e+00 2.0319922986
```

Statistical inference for logistic regression

One of the main goals of fitting models is to do prediction.

That means to see how well can the model predict $\pi(x)$ (in this case), for any particular value x of interest.

After fitting the model, $\hat{\pi}(x)$ is given by

$$\underbrace{\hat{\pi}(x)}_{\text{fitted value}} = \frac{\exp(\hat{\alpha} + \hat{\beta}x)}{1 + \exp(\hat{\alpha} + \hat{\beta}x)}.$$

We can use software to compute fitted values, as well as corresponding CIs.

Statistical inference for logistic regression

Below we compute the fitted values and CIs for all 173 observations of the *crab* data set:

```
# crab.fit <- glm(y ~ width, family = binomial, data   = crab.df)

# Fitted values of the **linear component (alpha + beta*x)**:
crab.predlin <- predict(crab.fit, type = "link", se.fit = TRUE)
crab.predlin$fit    # the fitted values
crab.predlin$se.fit # the SEs
```

Output:

```
# The fitted values:
         1           2           3          ...
 1.72080789 -1.16312951  0.57717754         ...

# The SEs:
        1         2         3       ...
0.3096620 0.3765148 0.1753939       ...
```

Statistical inference for logistic regression

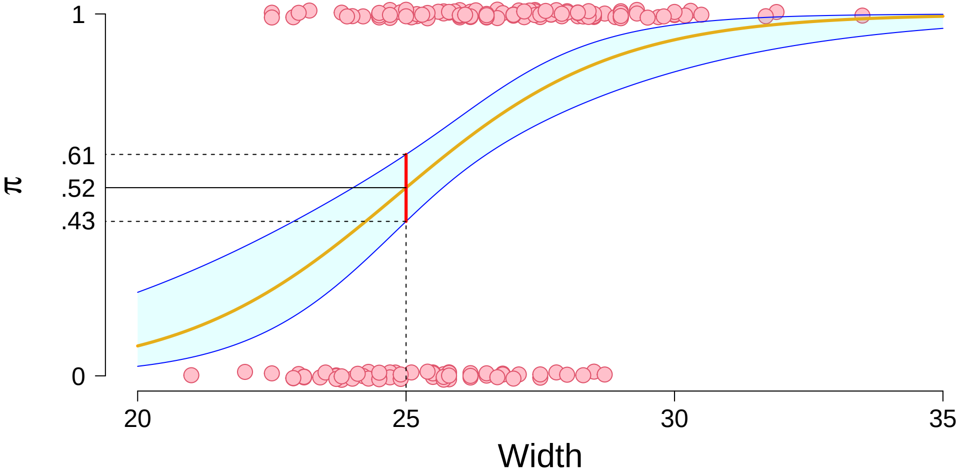
```r
# 95% CIs of the **linear component (alpha + beta*x)**:
predlinCI.LB <- crab.predlin$fit - (1.96 * crab.predlin$se.fit)
predlinCI.UB <- crab.predlin$fit + (1.96 * crab.predlin$se.fit)

# Fitted values and corresponding 95% CIs of **P(Y = 1)**:
fit.pi    <- plogis(crab.predlin$fit)
# same as exp(crab.predlin$fit) / (1 + exp(crab.predlin$fit))
predCI.LB <- plogis(predlinCI.LB)
predCI.UB <- plogis(predlinCI.UB)
# Summary:
cbind(width=crab.df$width, y = crab.df$y,
      predCI.LB, fit.pi, predCI.UB)
```

Output:

```
    width y  predCI.LB    fit.pi predCI.UB
1    28.3 1 0.75284998 0.8482329 0.9111490
2    22.5 0 0.12998419 0.2380991 0.3952826
3    26.0 1 0.55808789 0.6404177 0.7152356
------------------------------------------
```

Statistical inference for logistic regression



For instance, $\hat{\pi}(x = 25) = .52$, 95% prediction interval $=$ $(.43, .61)$.

Exercise 7-2

Refer to the previous exercise. Use the outputs obtained from the previous exercise.

a. Conduct a Wald test for the LI effect. Interpret.
b. Construct a Wald confidence interval for the odds ratio. Interpret.
c. Conduct a likelihood-ratio test for the LI effect. Interpret.
d. Construct the likelihood-ratio confidence interval for the odds ratio. Interpret.

Logistic regression with categorical predictors

Logistic regression with categorical predictors

Including categorical predictors (aka factors) in a logistic regression model is the same as with ordinary regression models:

One must use indicator (or 'dummy') variables.

Remember: A factor with k levels requires $(k-1)$ indicators.

Logistic regression with categorical predictors – Example

		Marijuana Use	
Race	Gender	Yes	No
White	Female	420	620
	Male	483	579
Other	Female	25	55
	Male	32	62

Predictors: *Race* and *Gender*.
Outcome: *Marijuana Use*, binary (0 = no; 1 = yes).

Both *Race* and *Gender* are factors with two levels.
Each therefore requires one indicator.

We will rely on R's default (indicator = last level in alphabetic order):

- *Race:* White.
- *Gender:* Male.

Logistic regression with categorical predictors – Example

```
# Marijuana data:
mar.use <- data.frame(Race   = c("white", "white", "other", "other"),
                      Gender = c("female", "male", "female", "male"),
                      Yes    = c(420, 483, 25, 32),
                      No     = c(620, 579, 55, 62))
# Fit logistic regression:
mar.use.fit <- glm(Yes / (Yes + No) ~ Race + Gender,
                   weights = Yes + No,
                   family  = binomial,
                   data    = mar.use)
summary(mar.use.fit)
```

Output:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.83035    0.16854  -4.927 8.37e-07 ***
racewhite    0.44374    0.16766   2.647  0.00813 **
gendermale   0.20261    0.08519   2.378  0.01739 *
```

Logistic regression with categorical predictors – Example

$$\log\left(\frac{\pi}{1-\pi}\right) = -0.83 + 0.20\text{Gender}_{\text{Male}} + 0.44\text{Race}_{\text{White}}.$$

Interpreting the effect of $\text{Gender}_{\text{Male}}$ on $\pi = P(Y = 1)$:

- *Additive:*
  Conditional on race (i.e., keeping race fixed), the estimated log odds that a male uses marijuana is 0.20 higher than the estimated log odds that a female uses marijuana.
- *Multiplicative:*
  Conditional on race, the estimated odds that a male uses marijuana is $e^{0.20} = 1.22$ times the estimated odds that a female uses marijuana.

Logistic regression with categorical predictors – Example

Statistical inference works exactly the same way (Wald, likelihood ratio).

Here's just one example:

Run a likelihood ratio test for $\mathcal{H}_0 : \beta_{\text{Gender}_{\text{Male}}} = 0$.

We just need to compare our model to a model that only includes predictor *Race*:

```r
# Fit logistic regression with Race only:
mar.use.fit2 <- glm(Yes / (Yes + No) ~ Race,
                    weights = Yes + No,
                    family  = binomial,
                    data    = mar.use
)

anova(mar.use.fit2, mar.use.fit,
      test = "LRT")
```

Logistic regression with categorical predictors – Example

Output:

```
Model 1: Yes/(Yes + No) ~ Race
Model 2: Yes/(Yes + No) ~ Race + Gender
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         2     5.7242
2         1     0.0580  1   5.6662   0.0173 *
```

$\chi^2(1) = 5.67$, $p = .017$: *Controlling for Race, we reject the hypothesis that Gender has no effect on marijuana use, at 5% significance level.*

Logistic regression with categorical predictors – Example

$$\operatorname{logit}(\pi) = -0.83 + 0.20\mathsf{Gender}_{\mathsf{Male}} + 0.44\mathsf{Race}_{\mathsf{White}}.$$

In terms of prediction, the model above makes one fixed prediction per (Gender, Race) combination:

Gender	Race	Prediction
female	other	$\hat{\pi} = \operatorname{logit}^{-1}(-0.83) = .30$
female	white	$\hat{\pi} = \operatorname{logit}^{-1}(-0.83 + 0.44) = .40$
male	other	$\hat{\pi} = \operatorname{logit}^{-1}(-0.83 + 0.20) = .35$
male	white	$\hat{\pi} = \operatorname{logit}^{-1}(-0.83 + 0.20 + 0.44) = .45$

Multiple logistic regression

Multiple logistic regression

Multiple $=$ more than one predictor.

$$\underbrace{\log\left(\frac{\pi}{1-\pi}\right)}_{\text{logit}(\pi)} = \alpha + \beta_1 x_1 \underbrace{+ \beta_2 x_2 + \cdots + \beta_p x_p}_{\text{more effects}}.$$

We've just seen one such model! It included two categorical predictors (one indicator each).

In general, predictors can be categorical and/or continuous.

Multiple logistic regression

Multiple $=$ more than one predictor.

$$\underbrace{\log\left(\frac{\pi}{1-\pi}\right)}_{\text{logit}(\pi)} = \alpha + \beta_1 x_1 \underbrace{+ \beta_2 x_2 + \cdots + \beta_p x_p}_{\text{more effects}}.$$

Interpret the effect of, say, x_1 on $\pi = P(Y = 1)$:

- *Additive:*
  Conditional on $x_2, \ldots, x_p$ (i.e., keeping them fixed), the estimated log odds is added β_1 units for each 1 unit increase of x_1.
- *Multiplicative:*
  Conditional on $x_2, \ldots, x_p$ (i.e., keeping them fixed), the estimated odds are multiplied by e^{β_1} for each 1 unit increase of x_1.

Multiple logistic regression - Example in R

width	color	y
28.3	2	1
22.5	3	0
26.0	1	1
⋮	⋮	⋮
28.0	1	0
27.0	4	0
24.5	2	0

173 rows in total.

Predictors:

- *width*, shell width in cm (continuous variable).
- *color*, shell color (categorical: 1 = medium light, 2 = medium, 3 = medium dark, 4 = dark). The darker, the older.

Outcome: y, binary (0 = female has no satellites; 1 = has satellites).

Multiple logistic regression – Example in R

Fit multiple logistic regression:

```r
crab.fit2 <- glm(y ~ width + factor(color),
                 family = binomial(link = "logit"),
                 data   = crab.df
                 )

summary(crab.fit2)
```

Output:

```
---------------------------
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -11.38519    2.87346  -3.962 7.43e-05 ***
width            0.46796    0.10554   4.434 9.26e-06 ***
factor(color)2   0.07242    0.73989   0.098    0.922
factor(color)3  -0.22380    0.77708  -0.288    0.773
factor(color)4  -1.32992    0.85252  -1.560    0.119
---------------------------
```

Multiple logistic regression – Example in R

```
---------------------------
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -11.38519    2.87346  -3.962 7.43e-05 ***
width            0.46796    0.10554   4.434 9.26e-06 ***
factor(color)2   0.07242    0.73989   0.098    0.922
factor(color)3  -0.22380    0.77708  -0.288    0.773
factor(color)4  -1.32992    0.85252  -1.560    0.119
---------------------------
```

By default, R created three coding variables for factor `color`:

Color	`factor(color)2`	`factor(color)3`	`factor(color)4`
1 = medium light	0	0	0
2 = medium	1	0	0
3 = medium dark	0	1	0
4 = dark	0	0	1

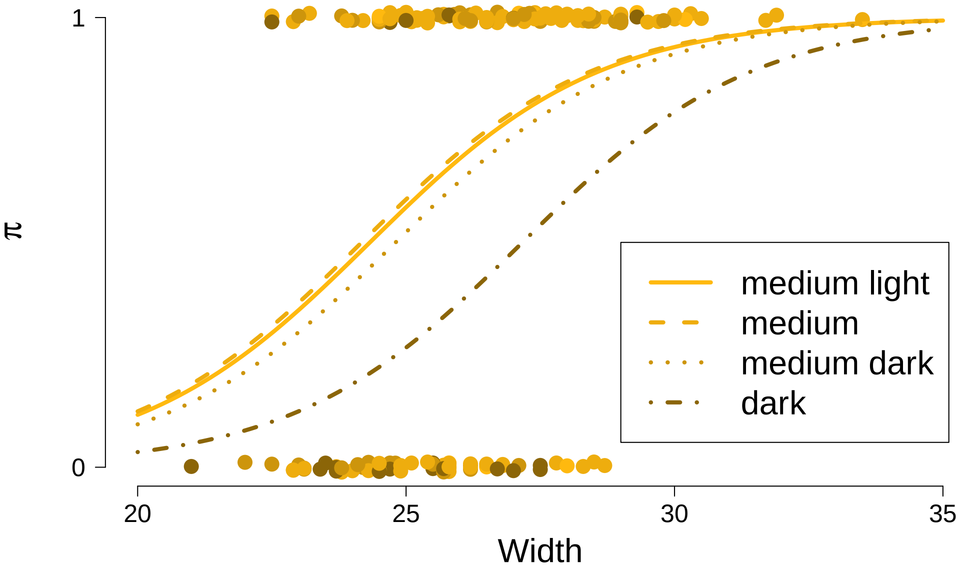
Multiple logistic regression – Example in R

$$\text{logit}(\hat{\pi}) = -11.385 + 0.468\text{width} + 0.072c_2 - 0.224c_3 - 1.330c_4.$$

Using the coding from the previous page, we find one model per color group:

Color	Prediction
1 = medium light	$\hat{\pi} = \text{logit}^{-1}(-11.385 + 0.468\text{width})$
2 = medium	$\hat{\pi} = \text{logit}^{-1}\left[(-11.385 + 0.072) + 0.468\text{width}\right]$
3 = medium dark	$\hat{\pi} = \text{logit}^{-1}\left[(-11.385 - 0.224) + 0.468\text{width}\right]$
4 = dark	$\hat{\pi} = \text{logit}^{-1}\left[(-11.385 - 1.330) + 0.468\text{width}\right]$

Multiple logistic regression – Example in R



Multiple logistic regression – Example in R

$$\text{logit}(\hat{\pi}) = -11.385 + 0.468\text{width} + 0.072c_2 - 0.224c_3 - 1.330c_4.$$

For all color groups, width $= 0.468$:

At each color group (i.e., keeping color fixed), the odds of having at least one satellite are multiplied by $e^{0.468} = 1.60$ *per 1cm increase of shell width.*

Multiple logistic regression – Example in R

$$\text{logit}(\hat{\pi}) = -11.385 + 0.468\text{width} + 0.072c_2 - 0.224c_3 - 1.330c_4.$$

At any fixed shell width, the odds of having at least one satellite are $e^{0.072} = 1.07$ *larger for color 2 (medium) than color 1 (medium light).*

Multiple logistic regression – Example in R

$$\text{logit}(\hat{\pi}) = -11.385 + 0.468\text{width} + 0.072c_2 - 0.224c_3 - 1.330c_4.$$

At any fixed shell width, the odds of having at least one satellite are $e^{-0.224} = .80$ *times smaller for color 3 (medium dark) than color 1 (medium light).*

Multiple logistic regression – Example in R

$$\text{logit}(\hat{\pi}) = -11.385 + 0.468\text{width} + 0.072c_2 - 0.224c_3 - 1.330c_4.$$

At any fixed shell width, the odds of having at least one satellite are $e^{-1.330} = .26$ *times smaller for color 4 (dark) than color 1 (medium light).*

Model comparison for nested models

Model comparison for nested models

We can proceed similarly as learned before, by comparing the deviances of the models of interest.

Model comparison for nested models - Example in R

- M_0: $\text{logit}(\hat{\pi}) = -12.35 + 0.50\text{width}$

```r
crab.fit <- glm(y ~ width,
                family = binomial(link = "logit"),
                data   = crab.df
                )

logLik(crab.fit)
deviance(crab.fit)
```

Output:

```
'log Lik.' -97.22633 (df=2)
[1] 194.4527
```

Model comparison for nested models - Example in R

- M_1: $\text{logit}(\hat{\pi}) = -11.39 + 0.47\text{width} + \underbrace{0.07c_2 - 0.22c_3 - 1.33c_4}_{\text{'color' effect}}$

```
crab.fit2 <- glm(y ~ width + factor(color),
                 family = binomial(link = "logit"),
                 data   = crab.df
                 )

logLik(crab.fit2)
deviance(crab.fit2)
```

Output:

```
'log Lik.' -93.72852 (df=5)
[1] 187.457
```

Model comparison for nested models - Example in R

Model	Log-lik	Deviance	df
M_0	$L_0 = -97.23$	$D_0 = 194.45$	2
M_1	$L_1 = -93.73$	$D_1 = 187.46$	5

$$D_0 - D_1 = 2(L_1 - L_0) = 6.99 \sim \chi^2(5-2) = \chi^2(3).$$

Assume $\alpha = 5\%$. Since

$$p\text{-value} = P(D_0 - D_1 > 6.99) = .07 > \alpha,$$

we conclude that the *color* effect does not significantly improve the model fit after controling for *width*.

Based on significance alone, M_0 is preferred.

Model comparison for nested models - Example in R

```r
anova(crab.fit, crab.fit2, test = "LRT")
```

Output:

```
Analysis of Deviance Table

Model 1: y ~ width
Model 2: y ~ width + factor(color)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       171     194.45
2       168     187.46  3   6.9956  0.07204 .
```

Interactions between explanatory variables

Interactions between explanatory variables

This is the same as in ordinary regression models.

For example:

$$\boxed{\text{logit}(\pi) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2}$$

Now the effect of x_1 on π depends on variable x_2:

$$\text{logit}(\pi) = \alpha + (\beta_1 + \beta_3 x_2) x_1 + \beta_2 x_2.$$

Similarly, the effect of x_2 on π depends on variable x_1:

$$\text{logit}(\pi) = \alpha + \beta_1 x_1 + (\beta_2 + \beta_3 x_1) x_2.$$

Interactions between explanatory variables – Example in R

```
crab.fit3 <- glm(y ~ width + factor(color) + width:factor(color),
                 family = binomial(link = "logit"),
                 data   = crab.df )
summary(crab.fit3)
```

Output:

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.75261   11.46409  -0.153    0.878
width                 0.10600    0.42656   0.248    0.804
factor(color)2       -8.28735   12.00363  -0.690    0.490
factor(color)3      -19.76545   13.34251  -1.481    0.139
factor(color)4       -4.10122   13.27532  -0.309    0.757
width:factor(color)2  0.31287    0.44794   0.698    0.485
width:factor(color)3  0.75237    0.50435   1.492    0.136
width:factor(color)4  0.09443    0.50042   0.189    0.850
----------------------------------------------------------
Residual deviance: 183.08  on 165  degrees of freedom
```

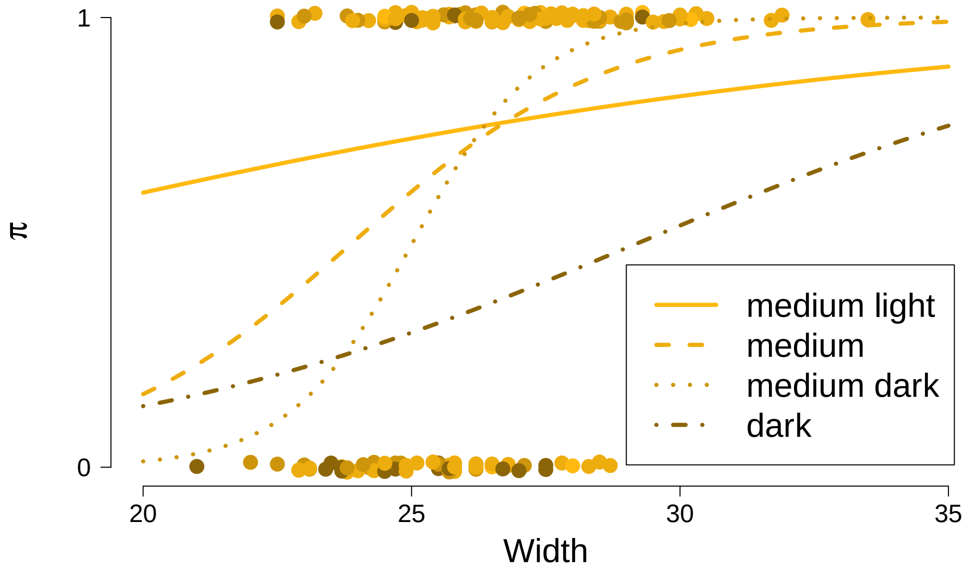
Interactions between explanatory variables – Example in R

$$\begin{aligned}\text{logit}(\hat{\pi}) = &-1.75 + 0.11\text{width} - 8.29c_2 - 19.77c_3 - 4.10c_4 \\ &+ 0.31(\text{width} \times c_2) + 0.75(\text{width} \times c_3) + 0.09(\text{width} \times c_4).\end{aligned}$$

Color	c_2	c_3	c_4
1 = medium light	0	0	0
2 = medium	1	0	0
3 = medium dark	0	1	0
4 = dark	0	0	1

Color	Prediction
1	$\hat{\pi} = \text{logit}^{-1}(-1.75 + 0.11\text{width})$
2	$\hat{\pi} = \text{logit}^{-1}\left[(-1.75 - 8.29) + (0.11 + 0.31)\text{width}\right]$
3	$\hat{\pi} = \text{logit}^{-1}\left[(-1.75 - 19.77) + (0.11 + 0.75)\text{width}\right]$
4	$\hat{\pi} = \text{logit}^{-1}\left[(-1.75 - 4.10) + (0.11 + 0.09)\text{width}\right]$

Interactions between explanatory variables – Example in R



Interactions between explanatory variables – Example in R

Was it worth to add the *width*-by-*color* interaction effect?

```
anova(crab.fit2, # width + factor(color)
      crab.fit3, # width * factor(color)
      test = "LRT")
```

Output:

```
Analysis of Deviance Table

Model 1: y ~ width + factor(color)
Model 2: y ~ width + factor(color) + width:factor(color)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       168     187.46
2       165     183.08  3   4.3764   0.2236
```

$\chi^2(3) = 4.38$, $p = .22$:
We conclude that, based on significance alone, we are better off not including the interaction effect.

Summarizing effects in logistic regression

Summarizing effects in logistic regression

We will see three ways of summarizing effects:

- Probability-based interpretations.
- Marginal effects and their average.
- Standardized interpretations.

Let's talk about each, one at a time.

Probability-based interpretations

To describe the effect of x_j on y, we can report $\hat{\pi} = \widehat{P}(Y = 1)$ while setting all remaining predictors at fixed, representative values (e.g., mean, quartiles, or using codes for code variables).

Here's an example.

Probability-based interpretations

$$\hat{\pi} = \text{logit}^{-1}\left[-11.385 + 0.468\text{width} + \underbrace{0.072c_2 - 0.224c_3 - 1.330c_4}_{\text{'color' effect}} \right]$$

The mean weight across color groups is 26.30 cm.

Let's compute the predicted probability of success for each color group, *at the mean weight*:

Color	Prediction
1 = medium light	$\hat{\pi} = \text{logit}^{-1}(-11.385 + 0.468 \times 26.30) = .72$
2 = medium	$\hat{\pi} = \text{logit}^{-1}[(-11.385 + 0.072) + 0.468 \times 26.30] = .73$
3 = medium dark	$\hat{\pi} = \text{logit}^{-1}[(-11.385 - 0.224) + 0.468 \times 26.30] = .67$
4 = dark	$\hat{\pi} = \text{logit}^{-1}[(-11.385 - 1.330) + 0.468 \times 26.30] = .40$

Thus, clearly, the darker the shell color, the lower the probability of satellites, *for crabs with a mean shell width*.

Probability-based interpretations – In R

```
predict(crab.fit2,
        newdata = data.frame(width = mean(crab.df$width), # 26.30
                             color = 1:4),                # the 4 colors
        type    = "response"                              # to get P(Y = 1)
        )
```

Output:

```
        1         2         3         4
0.7153494 0.7298626 0.6676801 0.3992933
```

Marginal effects and their average

Let $\mathbf{x} = (x_1, \dots, x_j, \dots, x_p)$.

As done before for the one-variable case, we can now also approximate $\pi(\mathbf{x})$ at some value x_j by the tangent at x_j, *while keeping the remaining predictors fixed*:

$$\pi'(\mathbf{x}) \simeq \beta_j \pi(\mathbf{x})[1 - \pi(\mathbf{x})].$$

This is called a marginal effect (i.e., across all other predictors).

Interpretation:

For values of the j-th predictor around x_j, a 1-unit increase in x_j corresponds approximately to a $\beta_j \pi(\mathbf{x})[1 - \pi(\mathbf{x})]$ change in $\pi(\mathbf{x})$.

Marginal effects and their average – Example

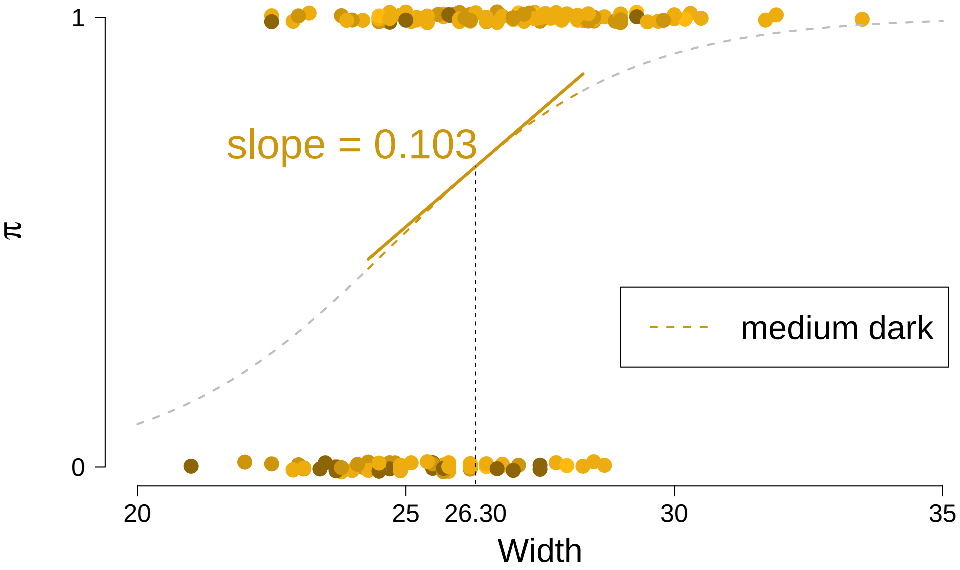
$$\text{logit}(\hat{\pi}) = -11.385 + 0.468\text{width} + \underbrace{0.072c_2 - 0.224c_3 - 1.330c_4}_{\text{'color' effect}}$$

We already saw that:

- The mean width in the sample is 26.30 cm.
- For the medium dark color group (color $= 3$), $\hat{\pi} = .67$.

Then, a 1-unit increase in *width* corresponds approximately to a $0.468 \times .67 \times (1 - .67) = 0.103$ change in $\hat{\pi}$.

Marginal effects and their average – Example



Marginal effects and their average – Example in R

```r
library(logitmfx)

logitmfx(crab.fit2, atmean = TRUE, data = crab.df)
```

Output:

```
---------------------------------------------------------
Marginal Effects:
                   dF/dx Std. Err.       z     P>|z|
width           0.102501  0.022194  4.6185 3.865e-06 ***
---------------------------------------------------------
```

Marginal effects and their average – Example in R

Instead of the *marginal* effect, we can also compute the average marginal effect.

It works like this:

- Compute the rate of change, $\beta\pi(1-\pi)$, for each observation in the sample.
- Average all these rates of change.

In R this is easy:

```
logitmfx(crab.fit2, atmean = FALSE, data = crab.df)
```

Output:

```
------------------------------------------------------------
Marginal Effects:
                     dF/dx Std. Err.        z    P>|z|
width             0.085312  0.024394   3.4973 0.00047 ***
------------------------------------------------------------
```

Marginal effects and their average – Example in R

$$\text{logit}(\hat{\pi}) = -11.385 + 0.468\text{width} + \underbrace{0.072c_2 - 0.224c_3 - 1.330c_4}_{\text{'color' effect}}$$

```
------------------------------------------------------------
Marginal Effects:
                   dF/dx Std. Err.       z   P>|z|
width           0.085312  0.024394  3.4973 0.00047 ***
------------------------------------------------------------
```

At the 173 observed width values, the average rate of change of $\hat{\pi}$ is 0.085 per 1-cm increase in width, adjusting for color.

Standardized interpretations

$$\boxed{\mathrm{logit}(\pi) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}$$

Just like ordinary multiple regression, we cannot directly compare regression effects β_i $(i = 1, \ldots, p)$ unless all predictors x_i are on the same units.

The way to avoid the problem is to standardize all predictors before fitting the model.

Thus, replace each predictor x_j by

$$\tilde{x}_j = \frac{x_j - \overline{x}_j}{s_{x_j}}.$$

Interpretation of $\tilde{\beta}_j$, the regression coefficient of $\tilde{x}_j$:

It is the change in π for each SD increase in x_j, adjusting for the other variables.

Summarizing predictive power in logistic regression

Summarizing predictive power in logistic regression

There are various ways of summarizing predictive power (i.e., how well the model predicts the response variable).

Here we will focus on only two:

- Classification tables.
- ROC curves.

Let's talk about each, one at a time.

Classification tables

A classification table cross-classifies *observed* with *predicted* Y values (0 or 1).

The problem is that logistic regression predicts $\pi = P(Y = 1)$, not Y itself.

To work around the issue, we need to transform probabilities π into 0-1 values.
This is done by using a cutoff value, π_0:

- If $\hat{\pi} \leq \pi_0$, then $\hat{y} = 0$;
- If $\hat{\pi} > \pi_0$, then $\hat{y} = 1$;

Classification tables

There are two natural choices for the cutoff π_0:

- $\pi_0 = .50$:
  OK as long as the observed y values are not dominated by 0s or by 1s.
- $\pi_0 = \overline{y}$:
  That is, the proportion of 1s in the sample.
  This is the better choice.

Classification tables – Example in R

```
crab.fit2 <- glm(y ~ width + factor(color),
                 family = binomial(link = "logit"),
                 data   = crab.df
                 )

pi_0    <- mean(crab.df$y)               # .64
pi.pred <- fitted(crab.fit2)              # predicted pi
y.pred  <- as.numeric(pi.pred > pi_0) # predicted y

# Classification table:
xtabs(~ crab.df$y + y.pred)
```

Output:

```
         y.pred
crab.df$y  0  1
        0 43 19
        1 36 75
```

Classification tables – Example in R

Output:

```
           y.pred
crab.df$y   0  1
          0 43 19
          1 36 75
```

The model correctly predicts 43 failures and 75 successes.

Some interesting quantities:

- Sensitivity $= P(\hat{y} = 1|y = 1) = \frac{75}{36+75} = .676.$
- Specificity $= P(\hat{y} = 0|y = 0) = \frac{43}{43+19} = .694.$
- Overall proportion of correct classifications $= \frac{43+75}{173} = .682.$

Classification tables

Classification tables, albeit handy, are not ideal:

- It enforces dichotomization of continuous $\hat{\pi}$ values.
- The choice of π_0 is rather arbitrary.
- Results depend on the proportion of 1s in response variable y.

ROC curves

ROC = receiver operating characteristic curve, plotting:

- y-axis: Sensitivity $= P(\hat{y} = 1 | y = 1)$

versus

- x-axis: $1-$ Specificity $= P(\hat{y} = 1 | y = 0)$,

for *all* possible cutoff values π_0 between 0 and 1.

The ROC curve is therefore more general than the classification table.

ROC curves

For $\pi_0 \simeq 0$ almost all $\hat{y}$ are 1, so...

- y-axis: Sensitivity $= P(\hat{y} = 1|y = 1) \simeq 1$
- x-axis: $1-$ Specificity $= P(\hat{y} = 1|y = 0) \simeq 1$.

For $\pi_0 \simeq 1$ almost all $\hat{y}$ are 0, so...

- y-axis: Sensitivity $= P(\hat{y} = 1|y = 1) \simeq 0$
- x-axis: $1-$ Specificity $= P(\hat{y} = 1|y = 0) \simeq 0$.

In general:
The ROC curve is nearly concave connecting the points $(0,0)$ (when $\hat{\pi} = 1$) and $(1,1)$ (when $\hat{\pi} = 0$).

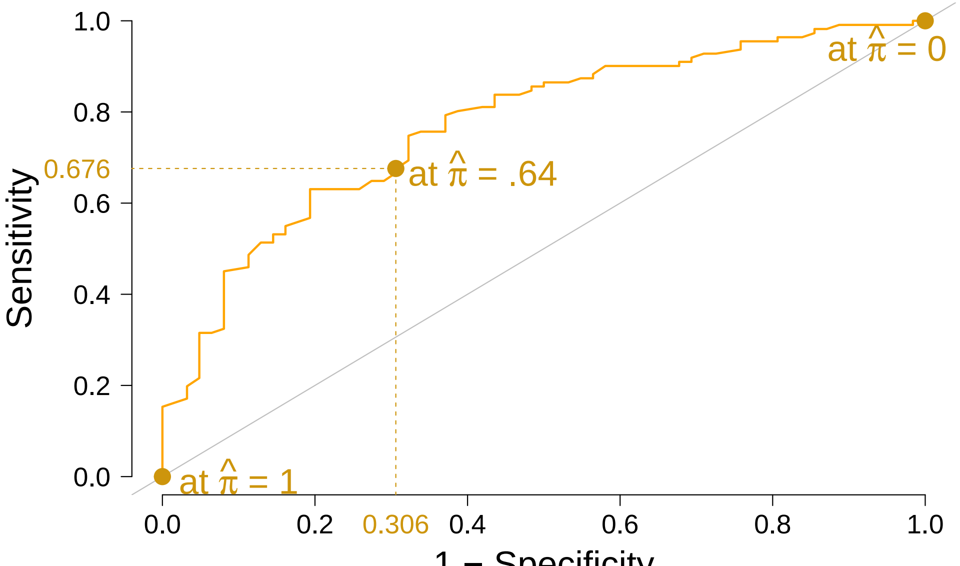
ROC curves – Example in R

```r
crab.fit2 <- glm(y ~ width + factor(color),
                 family = binomial(link = "logit"),
                 data   = crab.df
                 )

library(pROC)
crab.ROC <- roc(y ~ fitted(crab.fit2), data = crab.df)
plot.roc(crab.ROC,
         legacy.axes = TRUE # So that x-axis = 1 - specificity
         )
```

ROC curves – Example in R

```
## Warning: package 'pROC' was built under R version 4.4.2
```



ROC curves

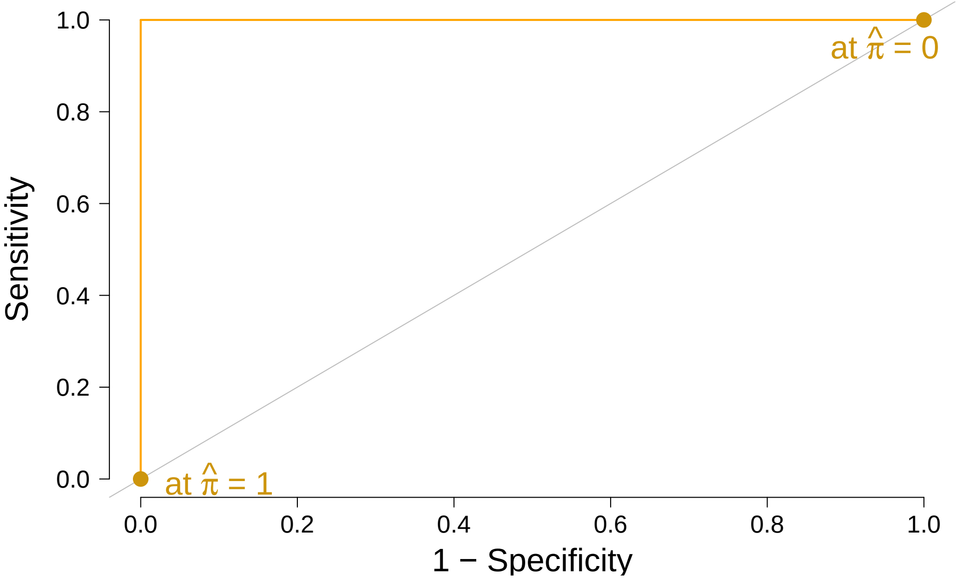
Ideally, for any x-value of the ROC curve,

$$y = \text{sensitivity} = P(\hat{y} = 1 | y = 1)$$

is as high as possible.

The idealized pattern would be as shown on the next page.

ROC curves



1.0
0.8
0.6
0.4
0.2
0.0
Sensitivity
at $\hat{\pi} = 0$
at $\hat{\pi} = 1$
0.0
0.2
0.4
0.6
0.8
1.0
1 − Specificity

ROC curves

The idealized pattern has an area under the curve (AUC) equal to 1.
Of course this is not possible with real data.

We typically compute and report the AUC, aka the concordance index.
The higher, the better.

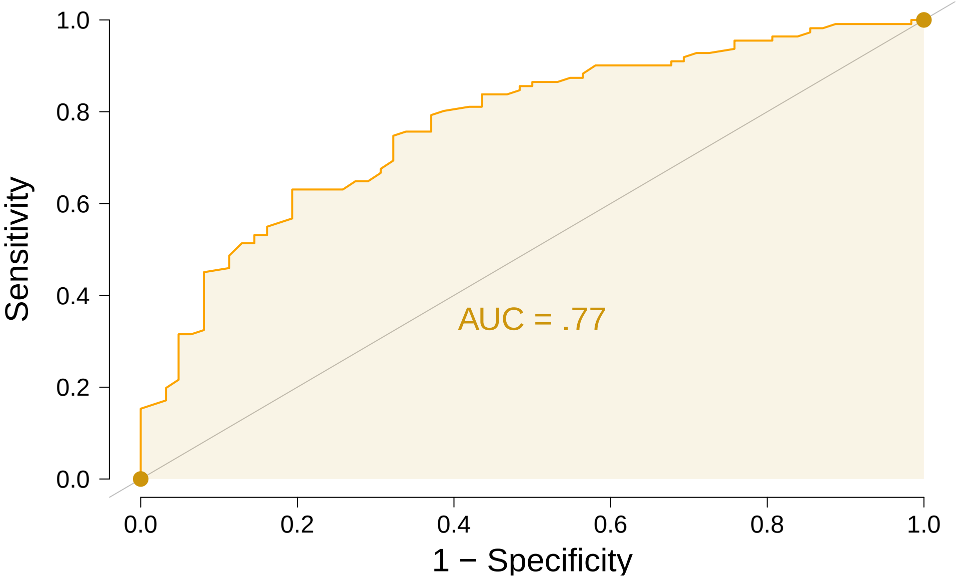
ROC curves - Example in R

```r
auc(crab.ROC)
```

Output:

```
Area under the curve: 0.7714
```

ROC curves – Example in R



Exercise 7-3

The below table shows the result of cross classifying a sample of people from the MBTI Step II National Sample (collected and compiled by CPP, Inc.) on whether they report drinking alcohol frequently (1 = Yes, 0 = No). There are four binary scales (categorical predictors) from a personality test: Extroversion/Introversion (E/I), Sensing/Intuitive (S/N), Thinking/Feeling (T/F) and Judging/Perceiving (J/P).

Extroversion/Introversion		E				I			
Sensing/iNtuitive		S		N		S		N	
		Alcohol Frequently							
Thinking/Feeling	Judging/Perceiving	Yes	No	Yes	No	Yes	No	Yes	No
T	J	10	67	3	20	17	123	1	12
	P	8	34	2	16	3	49	5	30
F	J	5	101	4	27	6	132	1	30
	P	7	72	15	65	4	102	6	73

Source: Reproduced with special permission of CPP Inc., Mountain View, CA 94043.

Exercise 7-3

Extroversion/Introversion		E				I			
Sensing/iNtuitive		S		N		S		N	
		Alcohol Frequently							
Thinking/Feeling	Judging/Perceiving	Yes	No	Yes	No	Yes	No	Yes	No
T	J	10	67	3	20	17	123	1	12
	P	8	34	2	16	3	49	5	30
F	J	5	101	4	27	6	132	1	30
	P	7	72	15	65	4	102	6	73

a. Import the dataset from file `MBTI_Ex7_3.dat` using the R command `read.table()` (note that the variable names are in the first row of the file). Save the dataset to object `MBTI`.

b. Fit model M_1 to these data, which should include the four scales as predictors of the probability of drinking alcohol frequently. Rely on R's default to create the required code variables. Report the estimated prediction equation for $\hat{\pi}$, explaining what each indicator variable stands for.

c. Based on M_1, compute $\hat{\pi}$ for a person of personality type (Extroversion, Sensing, Thinking, Judging).

Exercise 7-3

Extroversion/Introversion		E				I			
Sensing/iNtuitive		S		N		S		N	
		Alcohol Frequently							
Thinking/Feeling	Judging/Perceiving	Yes	No	Yes	No	Yes	No	Yes	No
T	J	10	67	3	20	17	123	1	12
	P	8	34	2	16	3	49	5	30
F	J	5	101	4	27	6	132	1	30
	P	7	72	15	65	4	102	6	73

Source: Reproduced with special permission of CPP Inc., Mountain View, CA 94043.

d. Based on the model parameter estimates, explain why the personality type with the highest $\hat{\pi}$ is (Extroversion, Intuitive, Thinking, Perceiving).

e. Interpret the effect of predictor TF in terms of the odds of $\pi = P(Y = 1)$.

f. Fit model M_2, including predictors EI and SN. Compare models M_1 and M_2 via a likelihood ratio test. What do you conclude?

Exercise 7-3

Extroversion/Introversion		E				I			
Sensing/iNtuitive		S		N		S		N	
		Alcohol Frequently							
Thinking/Feeling	Judging/Perceiving	Yes	No	Yes	No	Yes	No	Yes	No
T	J	10	67	3	20	17	123	1	12
	P	8	34	2	16	3	49	5	30
F	J	5	101	4	27	6	132	1	30
	P	7	72	15	65	4	102	6	73

g. Fit model M_3, including predictors EI, SN, and their interaction. Compare models M_2 and M_3 via a likelihood ratio test. What do you conclude?

h. What is the area under the curve (AUC) for model M_1?

Next lecture

In the next lecture, we cover Chapter 5.1, 5.2.

From the section above, I skipped: 5.1.4, 5.1.5, 5.2.3-5.2.5, 5.2.7-5.2.8.

Translated version of the exercises

There were some requests to have Japanese translated slides for the exercises. Therefore, the following slides provide Japanese version of the Exercises (for English students, the following pages are irrelevant).

Exercise 7-1 (日本語)

ある研究において, がん患者が寛解したかどうか (変数 $y = 1$ (yes), $y = 0$ (no)) と患者さんがトリチウム化チミジンを注射された後の細胞の増殖活性を測定するラベリングインデックス (LI = ラベル化された細胞の割合) の関連が調査された.

以下は取得したデータを分析する為のコードです. 以下のコードを R で実行し, 質問に答えなさい.

```
LI         <- c(8, 8, 10, 10, rep(c(12, 14, 16), each = 3),
                18, 20, 20, 20, 22, 22, 24, 26, 28, 32, 34, rep(38, 3))
y          <- c(rep(0, 13), 1, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0)
cancer.fit <- glm(y ~ LI, family = binomial)
summary(cancer.fit)
confint(cancer.fit)
```

a. LI = 26.0 の時, $P(Y = 1) = 0.50$ であることを示せ.

b. LI が 1 増加すると, 寛解の推定オッズが 1.16 倍になることを示せ.

c. LI の対数オッズを解釈し, 推定された LI の効果を解釈せよ.

d. $\hat{P}(Y = 1)$ の変化率は LI = 18 の時, 0.026 であることを示せ.

Exercise 7-2 (日本語)

Exercise 7-1 を参照し, そこから得られたアウトプットを使って以下の質問に答えよ.

a. LI の効果について Wald 検定を行い, 結果を解釈せよ.
b. オッズ比の Wald95% 信頼区間を求めて解釈せよ.
c. LI の効果について尤度比検定を行い, 結果を解釈せよ.
d. オッズ比の尤度比 95% 信頼区間を求めて解釈せよ.

Exercise 7-3 (日本語)

以下の表は MBTI の分類と飲酒の有無 (1 = Yes, 0 = No) の調査結果をまとめたものである. MBTI とは以下の 4 つの性格カテゴリーを用いて分類したものである: 外向性/内向性 (Extroversion/Introversion), 感覚的/直観的 (Sensing/iNtuitive), 思考的/感情的 (Thinking/Feeling), 判断/知覚 (Judging/Perceiving)

Extroversion/Introversion		E				I			
Sensing/iNtuitive		S		N		S		N	
		Alcohol Frequently							
Thinking/Feeling	Judging/Perceiving	Yes	No	Yes	No	Yes	No	Yes	No
T	J	10	67	3	20	17	123	1	12
	P	8	34	2	16	3	49	5	30
F	J	5	101	4	27	6	132	1	30
	P	7	72	15	65	4	102	6	73

Exercise 7-3 (日本語)

Extroversion/Introversion		E				I			
Sensing/iNtuitive		S		N		S		N	
		Alcohol Frequently							
Thinking/Feeling	Judging/Perceiving	Yes	No	Yes	No	Yes	No	Yes	No
T	J	10	67	3	20	17	123	1	12
	P	8	34	2	16	3	49	5	30
F	J	5	101	4	27	6	132	1	30
	P	7	72	15	65	4	102	6	73

a. Moodle の Example data から `MBTI_Ex7_3.dat` をダウンロードし, `read.table()` という R のコマンドを用いてデータを読み込む (ただし, 第 1 行目には変数名が入っているので注意). 読み込んだデータを `MBTI` というオブジェクト名に保存せよ. レポートにはその R コードを表示せよ.

b. 4 つの性格カテゴリーを説明変数とし, 飲酒の有無を予測したモデル M_1 を `MBTI` データに適用せよ. ただし, コード変数を作成する際に R のデフォルトを使用し, 推定結果を用いて $\hat{\pi}$ の予測式を表せ.

c. M_1 に基づくと, (Extroversion, Sensing, Thinking, Judging) の性格の人の $\hat{\pi}$ を求めよ.

Exercise 7-3

Extroversion/Introversion		E				I			
Sensing/iNtuitive		S		N		S		N	
		Alcohol Frequently							
Thinking/Feeling	Judging/Perceiving	Yes	No	Yes	No	Yes	No	Yes	No
T	J	10	67	3	20	17	123	1	12
	P	8	34	2	16	3	49	5	30
F	J	5	101	4	27	6	132	1	30
	P	7	72	15	65	4	102	6	73

d. モデルの推定結果に基づくと, $\hat{\pi}$ が最も高くなるのは (Extroversion, Intuitive, Thinking, Perceiving) という性格カテゴリーであることを説明せよ.

e. 説明変数 TF の効果を $\pi = P(Y = 1)$ のオッズの観点から解釈せよ.

f. EI と SN のみを説明変数としたモデルを M_2 とし, `MBTI` データに適用せよ. 尤度比検定を用いてモデル M_1 と M_2 を比較し, 解釈せよ.

Exercise 7-3

Extroversion/Introversion		E				I			
Sensing/iNtuitive		S		N		S		N	
		Alcohol Frequently							
Thinking/Feeling	Judging/Perceiving	Yes	No	Yes	No	Yes	No	Yes	No
T	J	10	67	3	20	17	123	1	12
	P	8	34	2	16	3	49	5	30
F	J	5	101	4	27	6	132	1	30
	P	7	72	15	65	4	102	6	73

g. EI と SN, およびこれらの交絡を説明変数としたモデルを M_3 とし, `MBTI` データに適用せよ. 尤度比検定を用いてモデル M_2 と M_3 を比較し, 解釈せよ.

h. M_1 モデルの are under the curve (AUC) を求めよ.