Contents

Logistic Regression

Suppose I want to predict a probability as a function of some independent variables, e.g., a yes-no response variable.

Linear regression problematic:

1. Probabilities are between 0 and 1; $\alpha + \beta x$ is unbounded
2. Residuals can take only two values — certainly not normally distributed
3. Variance of response, $\pi(1 - \pi)$, depends on the mean and is therefore heteroscedastic

Possible solutions:

1. Logistic regression
2. Discriminant analysis and naive/idiot Bayes classifiers

Outline of lecture:

1. The logistic regression model
2. Interpreting the results
3. Scoring other data sets

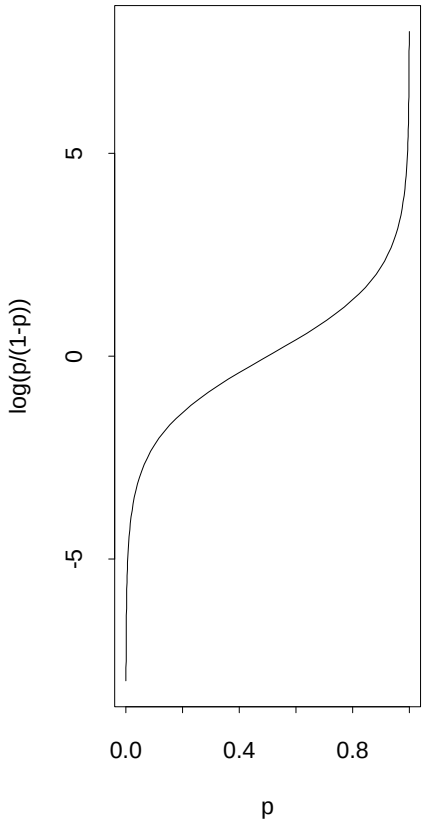
Logistic Regression Model

- Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a random sample from some population where $y_i \in \{0, 1\}$
- Let $\pi_i = \mathbb{E}(Y_i)$, i.e., probability person i responds "yes"

- We want to model π_i, but can't
- Instead, model **log-odds** or **logit** of π_i
- Odds $= \pi/(1-\pi)$
- Logistic regression:

$$\begin{aligned} \text{logit}(\pi_i) &= \log\left(\frac{\pi_i}{1-\pi_i}\right) \\ &= \alpha + \beta x_i \end{aligned}$$

- Estimate α and β with maximum likelihood



In general we have p predictors with

$$\text{logit}(\pi) = \alpha + \beta_1 x_1 + \cdots \beta_p x_p$$

Estimating Probabilities

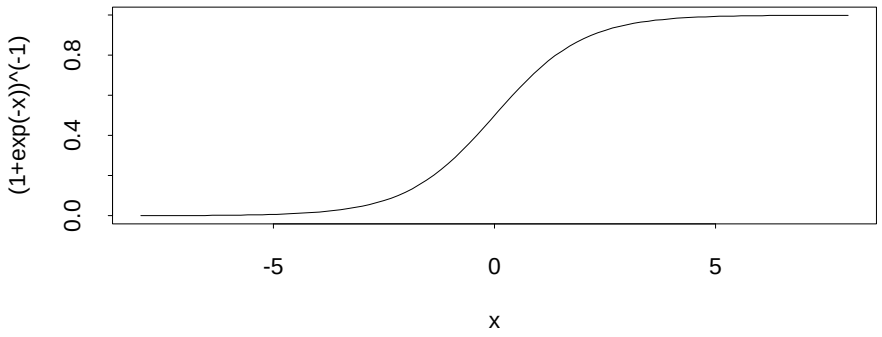
The logistic regression model is

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \alpha + \beta x_i$$

The log-odds ratio is interesting, but we often need probabilities at the end of the analysis, i.e., what is the probability of response? Answer: solve the above equation for π.
Let $\eta_i = \alpha + \beta x_i$

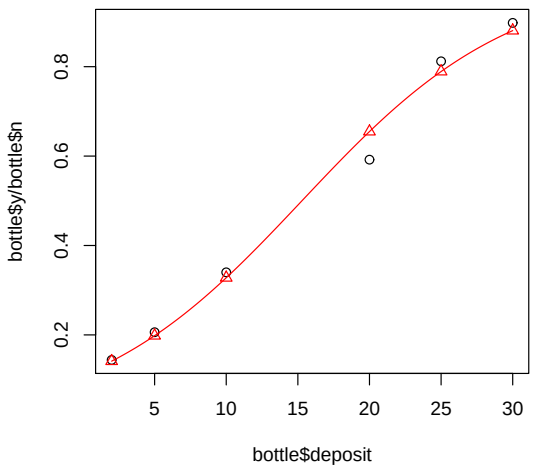
$$\frac{\hat{\pi}_i}{1-\hat{\pi}_i} = \exp(\eta_i)$$
$$\hat{\pi}_i = \exp(\eta_i) - \hat{\pi}_i \exp(\eta_i)$$
$$\hat{\pi}_i(1+\exp(\eta_i)) = \exp(\eta_i)$$
$$\hat{\pi}_i = \frac{\exp(\eta_i)}{(1+\exp(\eta_i))} = (1+\exp(-\eta_i))^{-1}$$

This is the **logistic function**. It's the most commonly used *squashing* or *sigmoidal* function used by neural network practitioners.



Bottle return problem

A carefully controlled experiment was conducted to study the effect of the size of the deposit on the likelihood that a returnable one-liter soft-drink bottle will be returned. A bottle return was coded 1 and no return was coded 0. The data show the number of bottles that were returned out of 500 sold at each of the six deposit levels. Plot estimated proportions against X. Estimate a logistic regression model and superimpose the fitted values. Interpret the parameter estimates.




```
bottle = data.frame(n=rep(500,6), deposit=c(2,5,10,20,25,30),y=c(72,103,170,296,406,449))
plot(bottle$deposit, bottle$y/bottle$n)
fit = glm(y/n ~ deposit, binomial, bottle, weight=n)
points(bottle$deposit, fit$fitted.values[1:6], pch=2, col=2)
x = seq(2,30,length=100)
lines(x, predict(fit, data.frame(deposit=x), type="response"), col=2)
summary(fit)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.076565   0.084839  -24.48   <2e-16 ***
x            0.135851   0.004772   28.47   <2e-16 ***
```

$$\log\left(\frac{\pi}{1-\pi}\right) = -2.08 + 0.136x$$

More bottle return problem

- Find a 95% confidence interval for β and test whether it is different from 0.

```
> confint(fit)
                 2.5 %     97.5 %
(Intercept) -2.2449682 -1.9123046
x            0.1266071  0.1453175
```

- What is the probability that a bottle will be returned when the deposit is 15 cents?

```
> eta = fit$coef[1] + fit$coef[2]*15  # linear predictor
> eta
-0.03880299
> predict(fit, data.frame(x=15))
-0.03880299

> 1/(1+exp(-(fit$coef[1] + fit$coef[2]*15))) # unlogit it
  0.4903005
> predict(fit, data.frame(x=15), type="response")
0.4903005
```

- For which deposit amount do you expect 75% of the bottles to be returned?

$$\log\left(\frac{.75}{1-.75}\right) = -2.08+0.136x \quad \Rightarrow \quad x = \frac{\log 3 + 2.08}{0.136} = 23.37$$

```
(log(3)-fit$coef[1])/fit$coef[2]
  23.37253
```

Logistic Regression Model

```
> fit = glm(2-q11 ~ food+atmosph+service, binomial, pizzarest)
> summary(fit)

Call:
glm(formula = 2 - q11 ~ food + atmosph + service, family = binomial,
    data = pizzarest)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.40245    0.40647  -5.911 3.41e-09 ***
food         1.38776    0.09934  13.970  < 2e-16 ***
atmosph     -0.13490    0.10159  -1.328    0.184
service      0.39515    0.09903   3.990 6.60e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2345.0  on 14732  degrees of freedom
Residual deviance: 2026.2  on 14729  degrees of freedom
  (1653 observations deleted due to missingness)
AIC: 2034.2

> vif(fit)
    food  atmosph  service
1.336162 1.454803 1.389792

> summary(glm(2-q11 ~ atmosph, binomial, pizzarest))

Call: glm(formula = 2 - q11 ~ atmosph, family = binomial, data = pizzarest)

            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.48454    0.29534   5.027 4.99e-07 ***
atmosph      0.68116    0.07832   8.697  < 2e-16 ***
```

Odds Ratio: e^{β}

```
> fit$coef
(Intercept)        food     atmosph     service
 -2.4024497   1.3877596  -0.1349032   0.3951547

> exp(fit$coef)
(Intercept)        food     atmosph     service
  0.0904960   4.0058652   0.8738005   1.4846138
```

What does the odds ratio mean? Consider two people:

1. `food` $= 4$, `atmosph` $= 4$, `service` $= 4$
2. `food` $= 5$, `atmosph` $= 4$, `service` $= 4$

The odds for person 1 are ($\pi_1 =$ prob person 1 says "yes")

$$\frac{\pi_1}{1-\pi_1} = \exp(\alpha + 4\beta_1 + 4\beta_2 + 4\beta_3)$$

The odds for person 2 are ($\pi_2 =$ prob person 2 says "yes")

$$\begin{aligned}
\frac{\pi_2}{1-\pi_2} &= \exp(\alpha + 5\beta_1 + 4\beta_2 + 4\beta_3) \\
&= \exp(\alpha + 4\beta_1 + \beta_1 + 4\beta_2 + 4\beta_3) \\
&= \exp(\alpha + 4\beta_1 + 4\beta_2 + 4\beta_3)e^{\beta_1} \\
&= \frac{\pi_1}{1-\pi_1}e^{\beta_1}
\end{aligned}$$

Thus, by increasing `food` by 1, the odds of saying "yes" are multiplied by e^{β_1}

Your Turn

1. In an experiment testing the effect of a toxic substance, 1,500 experimental insects were divided at random into six groups of 250 each. The insects in each group were exposed to a fixed dose of the toxic substance. A day later, each insect was observed. Death from exposure was scored 1 and survival was scored 0. The results are in the data frame below, where x_j is the dose level (on a logarithmic scale), administered to the insects in group j and y_j denotes the number of insects that dies out of the $n_j = 250$ in the group. As a hint, study the bottle return problem we worked in class.

```
toxicity = data.frame(x=1:6, n=rep(250,6), y=c(28,53,93,126,172,197))
fit = glm(y/n ~ x, binomial, toxicity, weight=n)
```

   (a) Plot the estimated proportions $p_j = y_j/n_j$ against x_j. Does the plot support the analyst's belief that the logistic response function is appropriate?

   (b) Find the MLEs of the slope and intercept, e.g., using `glm` in R. State the fitted response function and superimpose it on the scatterplot from part (a).

   (c) Obtain $\exp(b_1)$ and interpret this number.

   (d) What is the estimated probability that an insect dies when the dose level is $x = 3.5$?

   (e) What is the estimated median lethal dose—that is, the dose for which 50% of the experimental insects are expected to die?

   (f) Find a 99% confidence interval for β_1. Convert it into ones for the odds ratio.

2. The marketing manager for a large nationally franchised lawn service company would like to study the characteristics that differentiate home owners who do and do not have a lawn service. A random sample of 30 home owners located in a suburban area near a large city was selected. Predictor variables include household income ($K), lawn size (square feet K), attitude toward outdoor recreational activities (1=positive, 0=negative), number of teenagers in the household, and age of the head of household.

   (a) Generate a scatterplot matrix of the six variables. Comment on anything unusual or problematic.

   (b) Generate a correlation matrix of the six variables. Comment on substantial correlations.

   (c) Fit a logistic regression of whether a household has a lawn service (`lawnserv`) on the other five variables and test whether the overall regression model is significant.

   (d) State the estimated regression equation

   (e) Estimate the probability of purchasing a lawn service for a 45-year-old home owner with a family income of $70K, a lawn size of 3,000 square feet, a negative attitude towards outdoor recreation, and one teenager in the household.

(f) Which predictor variables are significantly different from zero (use alpha=.05)?

(g) Estimate a logistic regression model with income, attitude, and teenage as predictors. State the estimated regression equation and indicate which variables are significant.

(h) Estimate a logistic regression model with lawnsize, attitude, and teenage as predictors. State the estimated regression equation and indicate which variables are significant.

(i) Estimate a logistic regression model with HOH age, attitude, and teenage as predictors. State the estimated regression equation and indicate which variables are significant.

(j) Write a short paragraph with your final conclusions.

3. The data below give 4526 applicants (2691 males and 1835 females) who applied for admission to six departments in a university. The admission rate for males was 44.5% (1198/2691) and that for females was 30.4% (557/1835). This naturally raised the question of sex discrimination. Although the overall female admission rate was 14.1% lower for females than that for males, the female admission rate was actually higher than that for males in 4 out of 6 departments. This is called the *Simpson's paradox*. See here for more discussion.

```
dat = data.frame(female = c(rep(0,6), rep(1,6)), dept = rep(LETTERS[1:6],2),
  apps = c(825,560,325,417,191,373,108,25,593,375,393,341),
  admits = c(512,353,120,138,53,22,89,17,202,131,94,24))
```

(a) Explain why Simpson's paradox occurs for these data.

(b) Fir a logistic regression model to the data using gender as the only predictor. What does the slope tell you?

(c) Fit another logistic regression by adding department. Explain why the gender coefficient changes sign from positive to negative and how this illustrates Simpson's paradox.

Answers

1. (a) The plot looks like the logistic response is appropriate. (b) $\log[\pi/(1-\pi)] = -2.64 + 0.674x$. (c) We find $e^{0.674} = 1.96$, so for every additional log step in toxicity the odds of dying are roughly doubled. (d) 0.4293. (e) 2.644/0.6740 = 3.92. (f) Use `confint` to find .575, .777, then exponentiate it to find 1.78 to 2.18.

```
plot(toxicity$x, toxicity$y/toxicity$n) # part a
tox2 = data.frame(x=rep(1:6,2), y=c(rep(0,6), rep(1,6)),
  count=c(250-toxicity$y, toxicity$y))
fit = glm(y~x, binomial, tox2, weight=count)
summary(fit)  # parts b
x=seq(1, 6, by=.1)
lines(x, predict(fit, data.frame(x=x), type="resp"), col=2)
exp(coef(fit)[1])  # part c
predict(fit, data.frame(x=3.5), type="resp") # part d
exp(confint(fit, level=.99))  # part f
```

2. (a) Multicollinearity will be a problem. There are no outliers. (b) There are substantial correlations between age, income and lawn size. (c) $H_0 : \beta_1 = \cdots = \beta_5$ versus H_1 at least one $\beta_j \neq 0$. R reports deviances: 41.5888 − 9.7803 = 31.808, which has a chi-squared distribution with 5 df. $P \approx 0$, so we reject H_0 and conclude that at least one variable is significant. (d) $\log[\pi/(1-\pi)] = -70.5 + 0.287\texttt{income} + 1.06\texttt{lawnsize} - 12.7\texttt{attitude} - 0.200\texttt{teenager} + 1.08\texttt{age}$. (e) See code below: 0.7576026. (f) None of them are significant, although income is close ($P = .0598$). (g) $\log[\pi/(1-\pi)] = -14.5 + 0.160\texttt{income} - 2.19\texttt{attitude} + 0.216\texttt{teenager}$. Income is significant. (h) $\log[\pi/(1-\pi)] = -1.97 + 0.594\texttt{lawnsize} - 1.65\texttt{attitude} - 0.441\texttt{teenager}$. Lawn size is significant. (i) $\log[\pi/(1-\pi)] = -11.6 + 0.336\texttt{age} - 2.59\texttt{attitude} - 0.795\texttt{teenager}$. Age is significant. (j) Older people tend to have higher income and have larger lawns. These three variables are all indicators of being "established." The more established a person is, the more likely a person is to adopt lawn service.

```
> predict(fit, data.frame(age=45, income=70,
  lawn_siz=3, attitude=0, teenager=1), type="resp")
```

3. (a) A large number of women (the most frequent) apply to department C, which has the lowest admission rate for both sexes. The two most frequent departments for men are A and

B, which also have the highest admission rates for both sexes. When we average over the departments, it appears that men have a higher rate. (b) The log odds ratio is $\log[(1493/1198)/(1278/557)] = -0.6103524$, which equals the logistic regression coefficient for female below.

```
glm(admits/apps ~ female, binomial, dat, apps) # part b
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.22013    0.03879  -5.675 1.38e-08 ***
female      -0.61035    0.06389  -9.553  < 2e-16 ***
```

```
# part c
glm(admits/apps ~ dept + female, binomial, dat, apps)
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.58205    0.06899   8.436   <2e-16 ***
deptB       -0.04340    0.10984  -0.395    0.693
deptC       -1.26260    0.10663 -11.841   <2e-16 ***
deptD       -1.29461    0.10582 -12.234   <2e-16 ***
deptE       -1.73931    0.12611 -13.792   <2e-16 ***
deptF       -3.30648    0.16998 -19.452   <2e-16 ***
female       0.09987    0.08085   1.235    0.217
```

Generalized Linear Models

- We observe $(\mathbf{x}_i, y_i)$, $i = 1, \ldots n$ and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\mathsf{T}$
- Classical linear model:
$$y_i = \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i + \epsilon_i,$$
where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- Note that $\mathbb{E}(y_i) = \mu_i = \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i$
- The generalized linear model (GLM) involves two functions
$$g(\mu_i) = \eta_i = \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i$$
  - *Linear component*: $\eta_i = \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_i$
  - **Link function**: let univariate function g be monotonic and differentiable
  - *Random component*: y_i are independent and from an exponential family, which implies the variance of y_i depends on μ_i through a variance function $\text{var}(y_i) = \phi\mathbb{V}(\mu_i)$ where ϕ is called the *dispersion parameter*.
- The classical linear model assumes $g(\mu) = \mu$, the identity function, and y_i has a normal distribution
- Logistic regression assumes $g(\mu) = \log[\mu/(1-\mu)]$ and y_i is a Bernoulli trial ($\mathbb{V}(y) = \mu(1-\mu)$). Other possible links for binary responses include
  - Probit $g(\mu) = \Phi^{-1}(\mu)$, where Φ is the cumulative standard normal distribution
  - Complementary log-log: $g(\mu) = \log(-\log(1-\mu))$
- Other frequently-used distributions for y include Poisson and Gamma

Pizza Hut Data

A random sample of $n = 220$ consumers were surveyed to evaluate the effect of price on the purchase of a pizza from Pizza Hut. Subjects were asked to suppose that they were going to have a large 2-topping pizza delivered to their residence. They were asked to select from either Pizza Hut or another pizzeria of their choice. The price they would have to pay to get a Pizza Hut pizza differed from survey to survey. The dependent variable is whether or not a student selected Pizza Hut and independent variables are price and sex.

1. Fit a logistic regression model.
2. Test whether the overall model is significant.
3. State the estimated regression equation.
4. Interpret the meaning of the coefficients and odds ratios.
5. Test whether each variable is significant.
6. Predict the probability that a female student will select Pizza Hut if the price is \$8.99. Repeat this for prices of \$11.49 and \$13.99.
7. Regress purchase on price for males only. Note the regression equation.
8. Regress purchase on price for females only. Note the regression equation.
9. Fit a logistic regression model with different slopes for males and females. Test whether the slopes are equal.

Maximum Likelihood Estimation

- Until now we've been using least squares to estimate parameters, e.g., regression

$$\min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_{\alpha,\beta} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

- We've used SSE as the objective and to evaluate fit, e.g.,

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Least squares won't work with other methods such as logistic regression or latent class analysis
- Alternative approach:
  - Maximum likelihood estimation (MLE)
  - (-2 log likelihood) is generalization of SSE
- Goals for today:
  - What is estimation by maximum likelihood?
  - What is -2 log likelihood?

MLEs of Proportions

- Suppose we draw a random sample of size $n = 5$ from some population, send them an offer, and $x = 2$ respond. What is our best guess of the response probability?
- Basic stats answer: $p = 2/5 = .4$. Rationale: common sense
- Maximum likelihood answer: pick π so that the probability of observing 2 responses in 5 tries given π is maximized
  - Let $L(\pi)$ be the probability (likelihood) of observing the data if the probability of response is π

$$L(\pi) = \binom{5}{2}\pi^2(1-\pi)^3$$

  - Let $l(\pi) = \log L(\pi)$, called the *log-likelihood*

$$l(\pi) = \log(10) + 2\log(\pi) + 3\log(1-\pi)$$

$$\frac{dl(\pi)}{d\pi} = \frac{2}{\pi} - \frac{3}{1-\pi} = 0 \Longrightarrow \hat{\pi} = \frac{2}{5}$$

Guess (π)	$L(\pi)$	$l(\pi)$	$-2l(\pi)$
.3	.3087	−1.1754	2.3508
.35	.3364	−1.0894	2.1788
.4	.3456	−1.0625	2.1249
.45	.3369	−1.0879	2.1759
.5	.3125	−1.1632	2.3263

MLEs of Means

- Suppose we draw a random sample of size $n = 3$ from a normal population with unknown mean μ and known variance $\sigma^2 = 5$. The observed values are $x_1 = 4$, $x_2 = 5$, and $x_3 = 6$. What is the best guess of μ?
- Basic stats answer: $\bar{x} = (4 + 5 + 6)/3 = 5$. Rationale: common sense
- Maximum likelihood answer: pick μ so that the probability of observing the three values given μ is maximized

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$

$$L(\mu) = \prod_{i=1}^{3} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right]$$

$$\begin{aligned} l(\mu) &= \sum_{i=1}^{3}\left[-\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}(x_i-\mu)^2\right] \\ &= -k_1 - k_2\sum_{i=1}^{3}(x_i-\mu)^2 \end{aligned}$$

$$\frac{dl(\mu)}{d\mu} = 2k_2\sum_{i=1}^{3}(x_i-\mu) = 0$$

$$\Longrightarrow \mu = \frac{1}{3}\sum_{i=1}^{3} x_i$$

$$\text{Note } \frac{d^2 l(\mu)}{d\mu^2} = -6k_2 < 0$$

Likelihood Function for Logistic Regression

- Assume we have a random sample (observations independent, each have same probability of selection) and that we make two measurements on each observation (x_i, y_i), where y_i is a 0-1 variable and $i = 1, \ldots, n$
- Let $\pi_i = \mathsf{P}(y_i = 1)$, i.e., prob(person i says yes), and

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i$$

- Note that the probability distribution for person i is

$$f_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

- Since observations are independent, the probability distribution (likelihood) and log-likelihood for our sample is

$$f(y_1, \ldots, y_n) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$\begin{aligned}
\log(f) &= \sum_{i=1}^{n} [y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)] \\
&= \sum_{i=1}^{n} \left[y_i \log(\frac{\pi_i}{1-\pi_i}) + \log(1-\pi_i) \right] \\
&= \sum_{i=1}^{n} y_i(\alpha + \beta x_i) - \sum_{i=1}^{n} \log[1 + \exp(\alpha + \beta x_i)]
\end{aligned}$$

We maximize this with respect to α and β

Log-likelihood, deviance, AIC

- Log-likelihood:

$$l = \log(L) = \sum_{i=1}^{n} y_i(\alpha + \beta x_i) - \sum_{i=1}^{n} \log[1 + \exp(\alpha + \beta x_i)]$$

- For bottle return problem:

```
> bot2 = data.frame(x=rep(bottle$deposit,2), y=c(rep(0,6), rep(1,6)),
  count=c(500-bottle$y, bottle$y))
> eta = fit$coef[1]+fit$coef[2]*bot2$x
> round(eta,2)
 [1] -1.80 -1.40 -0.72  0.64  1.32  2.00 -1.80 -1.40 -0.72  0.64  1.32  2.00
> fit$y
 1  2  3  4  5  6  7  8  9 10 11 12
 0  0  0  0  0  0  1  1  1  1  1  1
> sum(bot2$count*fit$y*eta) - sum(bot2$count*log(1+exp(eta)))
[1] -1531.436
> logLik(fit)
'log Lik.' -1531.436 (df=2)
```

- We will usually use the *deviance*: $-2l$. Think of deviance as SSE, measuring how much is unexplained by the model[1]

```
> -2*logLik(fit)
[1] 3062.872
> deviance(fit)
[1] 3062.872
```

- Or we will use AIC = deviance + 2(number parameters), which penalizes the fit measure by $2p$ (R counts intercept as a parameter)

```
> AIC(fit)
[1] 3066.872
```

[1] Deviance is really $-2(l - l_s)$, where l is the log likelihood of the fitted model and l_s is the log likelihood of the saturated model, substituting y_i for π_i., For logistic regression $l_s = 0$.

Residual versus null deviance

```
> summary(fit)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.076565   0.084839  -24.48   <2e-16 ***
x            0.135851   0.004772   28.47   <2e-16 ***

    Null deviance: 4158.9  on 11  degrees of freedom
Residual deviance: 3062.9  on 10  degrees of freedom
AIC: 3066.9
```

- R reports AIC and the *residual deviance* for the *full model*, i.e., the one with all predictors in the model.
- It also reports the *null deviance*, which is the deviance of the intercept-only model:

  ```
  > fit.null = glm(y~1, bot2, family=binomial, weight=count)
  > deviance(fit.null)
  [1] 4158.862
  > fit$null.deviance
  [1] 4158.862
  ```

  Think of the null deviance as SST, measuring how much variation in Y is unexplained by the intercept model.
- The *difference* between them measures how much variation is explained by the model and plays the role of extra sums of squares.

  ```
  > fit$null.deviance - fit$deviance
  [1] 1095.99
  ```

- The *difference* has a chi-squared distribution and can test overall signifiance, H_0 : all $\beta_j = 0$.

  ```
  > 1-pchisq(fit$null.deviance - fit$deviance, 1)
  [1] 0
  ```

Likelihood-Ratio Test

- Consider the *full model* with $p + q$ predictors

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha+\beta_1 x_1+\cdots+\beta_p x_p+\beta_{p+1}x_{p+1}+\cdots+\beta_{p+q}x_{p+q}$$

- The *reduced model* has only p predictors, i.e., the last q predictors have been dropped.

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p$$

- Let D_1 be the deviance of the full model, and D_2 be the deviance of the reduced model.
- We can test $H_0 : \beta_{p+1} = \cdots = \beta_{p+q} = 0$ with the test statistic $D_2 - D_1$, which has a chi-square distributions with q degrees of freedom

Likelihood-Ratio Test For Single Parameters

How can we use the likelihood-ratio test to compare the fits of the two models:

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1\texttt{food} + \beta_2\texttt{atmosph} + \beta_3\texttt{service}$$

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1\texttt{food} + \beta_3\texttt{service}$$

i.e., does `atmosph` affect the response?

Answer:

- Let $-2l_3$ be the maximized log-likelihood for the three-predictor model (2026.183 from slide 7)
- Let $-2l_2$ be the maximized log-likelihood for the two-predictor model (2027.956)

Then $(-2l_2) - (-2l_3))$ has a chi-squared distribution with 1 degree of freedom.

```
> drop1(fit, test="Chisq")
Model:
2 - q11 ~ food + atmosph + service
        Df Deviance    AIC     LRT   Pr(Chi)
<none>       2026.2 2034.2
food     1   2207.7 2213.7 181.555 < 2.2e-16 ***
atmosph  1   2028.0 2034.0   1.773    0.1830
service  1   2041.6 2047.6  15.398  8.71e-05 ***

> d2=deviance(glm(2-q11~food+service,binomial,pizzarest,subset=(!is.na(atmosph))))
> d2
[1] 2027.956
> 1-pchisq(d2-deviance(fit),1)
[1] 0.1829724
```

LRT for Fitness Club Data

- Test whether payment type is significant, while controlling for log down payment and log use.

```
> fit = glm(default~log(downpmt+1)+log(use+1)+pmttype, binomial, default)
> summary(fit)
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        2.58488    0.08669   29.82   <2e-16 ***
log(downpmt + 1)  -0.69243    0.01822  -38.00   <2e-16 ***
log(use + 1)      -1.52831    0.04855  -31.48   <2e-16 ***
pmttypeStatement  -0.72340    0.05361  -13.49   <2e-16 ***
pmttypeCheck EFT  -3.99025    0.14181  -28.14   <2e-16 ***
pmttypeCredit EFT -2.94096    0.10685  -27.52   <2e-16 ***

    Null deviance: 17734  on 24842  degrees of freedom
Residual deviance: 11477  on 24837  degrees of freedom

> drop1(fit, test="Chisq")
Model:
default ~ log(downpmt + 1) + log(use + 1) + pmttype
                 Df Deviance   AIC    LRT   Pr(Chi)
<none>                11477 11489
log(downpmt + 1)  1    13236 13246 1758.9 < 2.2e-16 ***
log(use + 1)      1    12933 12943 1455.9 < 2.2e-16 ***
pmttype           3    14324 14330 2846.6 < 2.2e-16 ***
```

- Where does 2846.6 come from?

```
> fit2 = glm(default~log(downpmt+1)+log(use+1), binomial, default)
> deviance(fit2)-deviance(fit)
[1] 2846.583
> 1-pchisq(deviance(fit2)-deviance(fit), 3)
[1] 0
```

Predictive Accuracy of Classifiers

Suppose we are using a GLM (e.g., logit or probit) with p parameters to estimate a binary response

- GLMs minimize deviance, which has the same problems in detecting overfitting as SSE
- AIC = deviance + $2p$ is a commonly-used penalized measure
- Alternative measures include the *classification rate*, which is the percentage of correctly classified cases, and the *misclassification rate*, which is 1− classification rate. These should be computed on non-training data.
- **Confusion matrix**

	Predicted bad	**Predicted good**
Actual Bad	TN=(# True Negatives)	FP=(# False Positives)
Actual Good	FN=(# False Negatives)	TP=(# True Positives)

$$\textbf{Classification rate} = \textbf{accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{FP} + \text{FN} + \text{TP}}$$

- **Recall**: percent of relevant (good) items recommended, a.k.a. **true positive rate** (TPR) and **sensitivity**

$$\textbf{Recall} = \mathsf{P}(\text{predict good}|\text{actually good}) = \frac{\text{TP}}{\text{FN} + \text{TP}}$$

- **False positive rate**: percent of bad items recommended

$$\textbf{FPR} = \mathsf{P}(\text{predict good}|\text{actually bad}) = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

Decision-support metrics

- **Precision**: percent of recommended items relevant.

$$\textbf{Precision} = \mathsf{P}(\text{actually good}|\text{predict good}) = \frac{\text{TP}}{\text{FP} + \text{TP}}$$

- Additional complication: how do we classify observations? Let c be some cutoff and p_i be the predicted probably for observation i. Classify i as a "yes" when $p_i > c$ and "no" otherwise.
- The choice of c depends on misclassification costs. Depending on the situation, the "cost" of a false negative may be much greater than the cost of a false positive, e.g., airport security screening or disease detection versus spam filters.
- Examine precision and recall separately or their harmonic mean is the F_1 measure

$$F_1 = \frac{2}{1/R + 1/P} = 2\frac{P \times R}{P + R}$$

- Also common to plot precision versus recall varying c.
- *Receiver operating characteristic* (good article on ROC) curves plot TRP against FPR for different values of c. **AUC** is the area under the ROC curve.
  - AUC=0.5 means random guessing—model worthless
  - AUC=1 means "crystal-ball" perfect classification

Defaulting Customer Example

- Note that if we classify all observations as "no" then the classification rate is 88.3%. The classes are said to be highly **imbalanced**. This is a stupid classifier, but we must beat it!

```
> table(default$default)/length(default$default)
       0        1
0.882963 0.117037
```

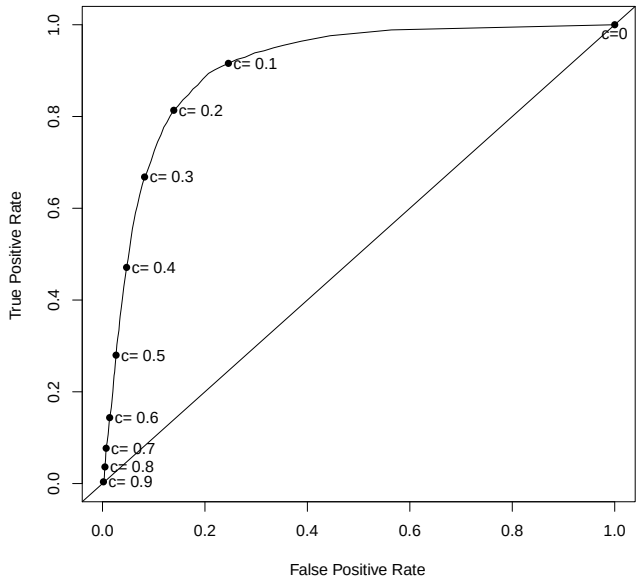
- Consider the following improved model

```
> fit = glm(default ~ log(downpmt+1)+pmttype+use+age+gender, binomial, default)
> tab=table(default$default, fit$fitted.values>.5)   # c=.5
> tab

    FALSE  TRUE
  0 21468   584
  1  2105   818
>
> sum(diag(tab))/sum(tab) # classification rate, (21468+818)/24975
[1] 0.8923323

> prop.table(tab,1)  # condition on observed values for FPR, TPR, etc.

         FALSE       TRUE
  0 0.97351714 0.02648286
  1 0.72015053 0.27984947

> tab=table(default$default, fit$fitted.values>.3) # c=.3
> sum(diag(tab))/sum(tab)
[1] 0.8886086
> prop.table(tab,1)  # TPR increases, but so does FPR

         FALSE       TRUE
  0 0.91783058 0.08216942
  1 0.33185084 0.66814916
```

How are FPR, TRP, etc. affected by class imbalance?

ROC Curves

```
a = (0:100)/100 # different cut points, don't call it c!
tpr = rep(NA, 101)  # true positive rate
fpr = rep(NA, 101)  # false positive rate
denom=table(default$default)
for(i in 1:101){
  num=table(default$default[fit$fitted.values>=a[i]])
  fpr[i] = num[1]/denom[1]
  tpr[i] = num[2]/denom[2]
}
plot(fpr, tpr, type="l", xlab="False Positive Rate", ylab="True Positive Rate")
abline(0,1)
b = (0:10)*10+1
points(fpr[b], tpr[b], pch=16)
text(fpr[b[-1]]+.01, tpr[b[-1]], paste("c=",as.character(a[b[-1]])), adj=0)
text(1,.98, "c=0")

library(pROC)
plot.roc(default$default, fit$fitted.values)
```

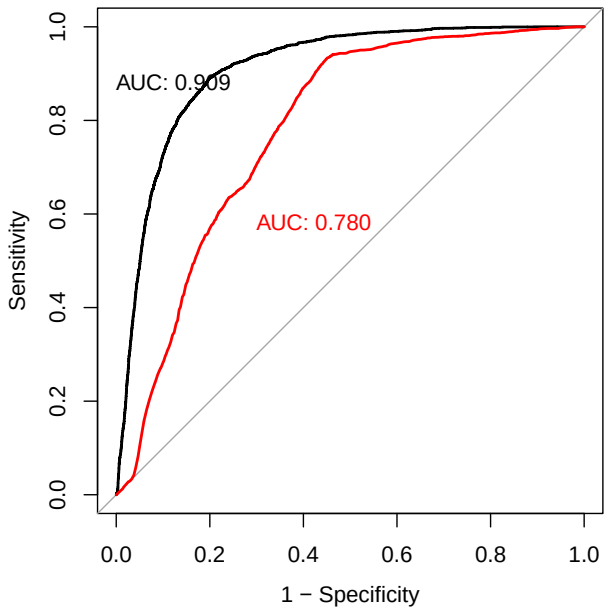


Area Under the ROC Curve (AUC) is used to evaulate models

ROC Curves

We are usually interested in comparing the ROC curves and AUC values for different models:

```
library(pROC)
ok = !is.na(default$use) # use has missing values
fit = glm(default ~ log(downpmt+1)+pmttype+use+age+gender, binomial, default, subset=ok)
plot.roc(default$default[ok], fit$fitted.values, legacy.axes=T,
    print.auc=T, print.auc.x=1, print.auc.y=.9)
fit2 = glm(default ~ pmttype+age+gender, binomial, default, subset=ok)
plot.roc(default$default[ok], fit2$fitted.values, add=T, col=2,
    print.auc=T, print.auc.x=.7, print.auc.y=.6, print.auc.col=2)
```



Important: ROC/AUC does not depend on the fraction of cases in each class (class imbalance problem), while classification rate does!

Your Turn

1. Use the estimates from the toxicity problem in the Your Turn on page 9 to generate an ROC curve and find the area under it. You have summarized data and I would like for you to generate the ROC curve "by hand." Here are hints.

   The are six values of $x = 1, \ldots, 6$ and let $\hat{\pi}_x$ be the predicted probability for x using the logistic regression model.

   (a) Complete the table to the right.

   (b) Plot TPR against FPR and find the area assuming a trapezoid between successive values.

   (c) Plot precision against recall.

Cut value	TPR	FPR
$0 \le c < \hat{\pi}_1$		
$\hat{\pi}_1 \le c < \hat{\pi}_2$		
$\hat{\pi}_2 \le c < \hat{\pi}_3$		
$\hat{\pi}_3 \le c < \hat{\pi}_4$		
$\hat{\pi}_4 \le c < \hat{\pi}_5$		
$\hat{\pi}_5 \le c < \hat{\pi}_6$		
$\hat{\pi}_6 \le c \le 1$		

Answers

(a) I have computed it exactly below. Note that there are 669 positives and $1500 - 669 = 831$ negatives, which provide the denominators for the TPR and FPR columns.

Cut value	# yes (Cum)	TPR	FPR	Area
$0 \le c < 0.1226$	0 (0)	$1 - \frac{0}{669} = 1$	$1 - \frac{0}{831} = 1$	
$0.1226 \le c < 0.2149$	28 (28)	$1 - \frac{28}{669} = 0.9581$	$1 - \frac{222}{831} = 0.7329$	0.2616
$0.2149 \le c < 0.3489$	53 (81)	$1 - \frac{89}{669} = 0.8789$	$1 - \frac{419}{831} = 0.4958$	0.2178
$0.3489 \le c < 0.5120$	93 (174)	$1 - \frac{174}{669} = 0.7399$	$1 - \frac{576}{831} = 0.3069$	0.1529
$0.5120 \le c < 0.6726$	126 (300)	$1 - \frac{300}{669} = 0.5516$	$1 - \frac{700}{831} = 0.1576$	0.0964
$0.6726 \le c < 0.8009$	172 (472)	$1 - \frac{472}{669} = 0.2945$	$1 - \frac{778}{831} = 0.0638$	0.0397
$0.8009 \le c \le 1$	197 (197)	$1 - \frac{669}{669} = 0$	$1 - \frac{831}{831} = 0$	0.0094
Total	669			0.77768

(b) See table above for .77768. For example, $.2616 = \frac{1}{2}(1 + .9581)(1 - .7329)$. Here is my R code.

```
toxicity=data.frame(x=1:6,n=rep(250,6),y=c(28,53,93,126,172,197))
toxlong = data.frame(
  x = c(rep(1,250),rep(2,250),rep(3,250),rep(4,250),
    rep(5,250),rep(6,250)),
  y = c(
    rep(1, 28), rep(0, 250-28), rep(1, 53), rep(0, 250-53),
    rep(1, 93), rep(0, 250-93), rep(1, 126), rep(0, 250-126),
    rep(1, 172), rep(0, 250-172), rep(1, 197), rep(0, 250-197)
  )
)
fit = glm(y/n ~ x, binomial, toxicity, weight=n)
toxicity$phat = fit$fitted.values
toxicity
  x   n   y      phat
1 1 250  28 0.1224230
2 2 250  53 0.2148914
3 3 250  93 0.3493957
4 4 250 126 0.5130710
5 5 250 172 0.6739903
6 6 250 197 0.8022286
fit2 = glm(y~x, binomial, toxlong)
summary(fit)
summary(fit2)
```

```
library(pROC)
plot.roc(toxlong$y, fit2$fitted.values, print.auc=T)
roc(toxlong$y, fit2$fitted.values)

myroc = data.frame(
  tpr=c(1,1-28/669,1-89/669,1-174/669,1-300/669,1-472/669,0),
  fpr=c(1,1-222/831,1-419/831,1-576/831,1-700/831,1-778/831,0)
)
plot(myroc$fpr, myroc$tpr, type="l")
```

To see how the numbers are computed, let's assume a cutoff of $c = .3$. Then the first two rows, which have 500 cases, are classified as no (see predicted probabilities in output). The remaining rows are classified by the model as yes, with 1000 cases. This would be the confusion matrix:

Truth	Model Prediction		Total
	False	True	
False	$500 - 81 = 419$	$831 - 419 = 412$	831
True	28+53=81	$669 - 81 = 588$	669
Total	500	1000	1500

Then TPR = 588/669 = 0.834 and FPR = 412/831 = 0.496. Repeate this for other cutoffs.