STP 530
Lecture 11: Logistic Regression

Yi Zheng, Ph.D.
yi.isabel.zheng@asu.edu

ARIZONA STATE UNIVERSITY
SCHOOL OF MATHEMATICAL AND STATISTICAL SCIENCES

November 7, 2023

Outline

1. The problem
2. The logistic regression model
3. The generalized linear models (GLM)
4. Making prediction with the fitted model
5. Interpreting model coefficients
6. Estimating model coefficients
7. Inferences
8. Model utility (practical significance) – pseudo R^2
9. Diagnostics
10. From predicted probability to binary classification

The problem: The *response variable* is binary.

Examples:

- Voting result between two candidates: Candidate A or Candidate B.
- Medical test result: positive or negative.
- Functional test of a product: success or failure.
- Survey response to a yes-no question.

Consider the regular multiple regression model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon$$

With the assumption $\varepsilon \; iid \sim N(0, \sigma^2)$, we have

$$\begin{aligned} E\{Y\} &= E\{\beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon\} \\ &= E\{\beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}\} + E\{\varepsilon\} \\ &= (\beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}) + (0) \\ &= \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} \end{aligned}$$

Regular multiple regression models use a linear combination of the predictors to predict $E\{Y\}$, the expected value of the response variable value.

Question: What is $E\{Y\}$ when Y is binary?

We usually model a binary variable with the Bernoulli distribution:

$$Y = \begin{cases} 1, & \text{with probability } P(Y=1) = \pi \\ 0, & \text{with probability } P(Y=0) = 1-\pi \end{cases}$$

The expected value and the variance of the distribution are:

$$E\{Y\} = \pi$$

$$\sigma^2\{Y\} = \pi(1-\pi)$$

Proof of $E\{Y\}$ and $\sigma^2\{Y\}$:

$$
\begin{aligned}
E\{Y\} &= \sum Y \cdot P(Y) \\
&= 1 \cdot \pi + 0 \cdot (1 - \pi) = \pi
\end{aligned}
$$

$$
\begin{aligned}
\sigma^2\{Y\} &= E\{(Y - E\{Y\})^2\} \\
&= \sum (Y - \pi)^2 \cdot P(Y) \\
&= (1 - \pi)^2 \pi + (0 - \pi)^2 (1 - \pi) \\
&= (1 - \pi)(\pi - \pi^2 + \pi^2) \\
&= \pi(1 - \pi)
\end{aligned}
$$

When Y is modeled with a Bernoulli distribution with $E\{Y\} = \pi$, the regular multiple regression model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon$$

implies

$$\pi = E\{Y\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

Is this model appropriate?

What's wrong with this model for a binary Y?

$$\pi = E\{Y\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

The issue:

π is the probability of success in the Bernoulli distribution, so $0 \leq \pi \leq 1$.

If ordinary linear models are used, there is nothing that restricts the right hand side of the equation within the $[0, 1]$ interval to match the domain of the left hand side.

The solution: Transform Y

Step 1: Take the odds

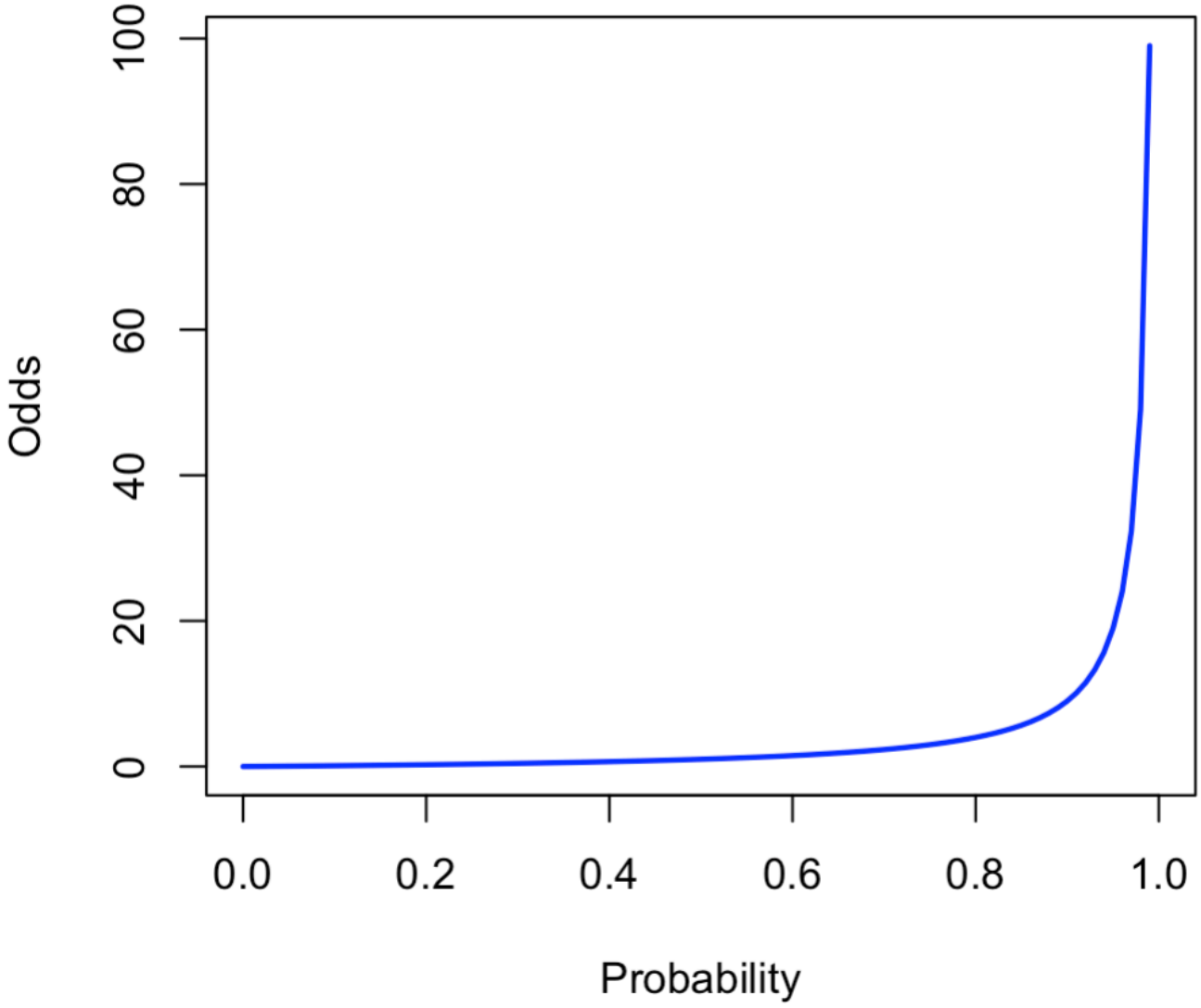
For a binary variable Y,

$$\text{odds} = \frac{P(\text{success})}{P(\text{failure})} = \frac{P(Y=1)}{P(Y=0)} = \frac{\pi}{1-\pi} = \frac{E\{Y\}}{1-E\{Y\}}$$

Compared with π, the odds is no longer bounded by 1 — it can be as large as positive infinity.

However, odds is still bounded by 0 — still does not match the unbounded linear predictor $\mathbf{X}_i\mathbf{b}$.

Keep in mind: The odds is on a different scale than probabilities.



Two steps of transformation

Step 2: Take the log

$$\text{logit} = \text{log-odds} = \ln \frac{\pi}{1-\pi} = \ln \frac{E\{Y\}}{1-E\{Y\}}$$

Logit is unbounded and matches the linear predictor $\mathbf{X}_i\mathbf{b}$.

Outline

1. The problem
2. The logistic regression model
3. The generalized linear models (GLM)
4. Making prediction with the fitted model
5. Interpreting model coefficients
6. Estimating model coefficients
7. Inferences
8. Model utility (practical significance) – pseudo R^2
9. Diagnostics
10. From predicted probability to binary classification

The logistic regression model

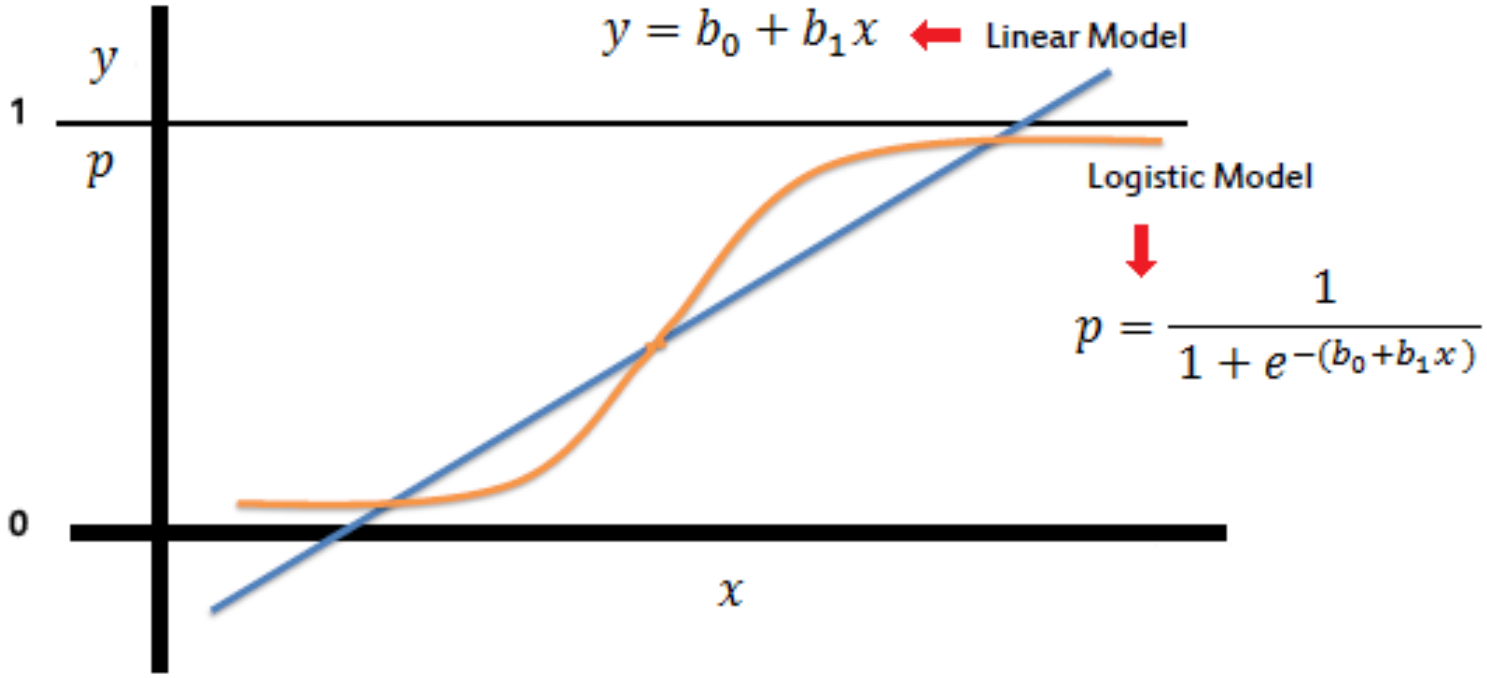
The original form:

$$\ln \frac{E\{Y\}}{1 - E\{Y\}} = \ln \frac{\pi}{1 - \pi} = \mathbf{X}\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

The operational form:

$$E\{Y\} = \pi = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1 + \exp(\mathbf{X}\boldsymbol{\beta})}$$

The logistic curve looks like an "S" shape. We can see how it constraints the expected value of Y within $(0, 1)$, as compared with ordinary linear regression.



Outline

1. The problem
2. The logistic regression model
3. The generalized linear models (GLM)
4. Making prediction with the fitted model
5. Interpreting model coefficients
6. Estimating model coefficients
7. Inferences
8. Model utility (practical significance) – pseudo R^2
9. Diagnostics
10. From predicted probability to binary classification

Logistic regression belongs to the family of **generalized linear models (GLM)**.

In GLM, the mean of the Y distribution $\mu = E\{Y\}$ depends on the predictors through a link function:

$$g(E\{Y\}) = \mathbf{X}\boldsymbol{\beta}$$

where $E\{Y\}$ is the expected value of Y given X's, $\mathbf{X}\boldsymbol{\beta}$ is called the **linear predictor**, a linear combination of the X values, and g is the **link function**.

Common GLMs:

1. The identity link $g(E\{Y\}) = E\{Y\}$; $Y|\mathbf{X}$ is assumed to follow a normal distribution; equivalent with OLS; In R, family = gaussian.
2. The log link $g(E\{Y\}) = \ln(E\{Y\})$; $Y|\mathbf{X}$ is assumed to follow a Poisson distribution; a.k.a., Poisson regression; In R, family = poisson.
3. The logit link $g(E\{Y\}) = \ln \frac{E\{Y\}}{1-E\{Y\}}$; $Y|\mathbf{X}$ is assumed to follow a Bernoulli distribution; a.k.a., logistic regression; In R, family = binomial.

Errors in GLM

The error term is **not additive** to the linear predictor in GLMs. The prediction error is modeled by the distribution of Y given X's, which is typically a function of the mean.

Diagnostics in GLM

Within the GLM framework, we no long need to check the distribution of the error term (i.e., homoskedasticity, normal distribution).

The other diagnostics are still applicable (e.g., multicollinearity, influential points, independent observations).

Outline

1. The problem
2. The logistic regression model
3. The generalized linear models (GLM)
4. Making prediction with the fitted model
5. Interpreting model coefficients
6. Estimating model coefficients
7. Inferences
8. Model utility (practical significance) – pseudo R^2
9. Diagnostics
10. From predicted probability to binary classification

Example: Gender discrimination in hiring

Suppose you are investigating allegations of gender discrimination in the hiring practices of a particular firm. An equal-rights group claims that females are less likely to be hired than males with the same background, experience, and other qualifications. You collected data on 28 former applicants.

$$Y = \begin{cases} 1, & \text{if hired} \\ 0, & \text{if not} \end{cases}$$

$$X_1 = \text{Years of higher education}$$

$$X_2 = \text{Years of experience}$$

$$X_3 = \begin{cases} 1, & \text{if male} \\ 0, & \text{if female} \end{cases}$$

The data

Hire (y)	Education (x_1)	Experience (x_2)	Gender (x_3)	Hire (y)	Education (x_1)	Experience (x_2)	Gender (x_3)
0	6	2	0	1	4	5	1
0	4	0	1	0	6	4	0
1	6	6	1	0	8	0	1
1	6	3	1	1	6	1	1
0	4	1	0	0	4	7	0
1	8	3	0	0	4	1	1
0	4	2	1	0	4	5	0
0	4	4	0	0	6	0	1
0	6	1	0	1	8	5	1
1	8	10	0	0	4	9	0
0	4	2	1	0	8	1	0
0	8	5	0	0	6	1	1
0	4	2	0	1	4	10	1
0	6	7	0	1	6	12	0

R Output

```
> m <- glm(HIRE~EDUC+EXP+GENDER, data=hire, family=binomial)
> summary(m)

Call:
glm(formula = HIRE ~ EDUC + EXP + GENDER, family = binomial,
    data = hire)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4380  -0.4573  -0.1009   0.1294   2.1804

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.2483     6.0805  -2.343   0.0191 *
EDUC          1.1549     0.6023   1.917   0.0552 .
EXP           0.9098     0.4293   2.119   0.0341 *
GENDER        5.6037     2.6028   2.153   0.0313 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35.165  on 27  degrees of freedom
Residual deviance: 14.735  on 24  degrees of freedom
AIC: 22.735
```

The fitted model

The original form:

$$\ln \frac{\pi}{1-\pi} = -14.25 + 1.15(\text{EDU}) + .91(\text{EXP}) + 5.6(\text{GENDERmale})$$

The operational form:

$$\begin{aligned}\pi =& \frac{\exp(-14.25 + 1.15(\text{EDU}) + .91(\text{EXP}) + 5.6(\text{GENDERmale})}{1 + \exp(-14.25 + 1.15(\text{EDU}) + .91(\text{EXP}) + 5.6(\text{GENDERmale})} \\ =& \frac{1}{1 + \exp[-(-14.25 + 1.15(\text{EDU}) + .91(\text{EXP}) + 5.6(\text{GENDERmale})]}\end{aligned}$$

where π is the probability that a job applicant gets hired.

Making prediction for existing cases in the dataset

```
> predict(m, type="link")
         1          2          3          4          5          6
-5.4992779 -4.0250566  3.7437960  1.0142501 -8.7188873 -2.2796685
         7          8          9         10         11         12
-2.2053593 -5.9893414 -6.4091265  4.0892720 -2.2053593 -0.4599712
        13         14         15         16         17         18
-7.8090387 -0.9500347  0.5241866 -3.6795806  0.5944650 -0.8054471
        19         20         21         22         23         24
-3.2597955 -3.1152079 -5.0794928 -1.7152958  5.1437082 -1.4400982
        25         26         27         28
-4.0993657 -0.8054471  5.0734298  3.5992085


> predict(m, type="response")
           1            2            3            4            5
0.0040730658 0.0175489472 0.9768829419 0.7338510832 0.0001634423
           6            7            8            9           10
0.0928208655 0.0992702542 0.0024990525 0.0016437556 0.9835245625
          11           12           13           14           15
0.0992702542 0.3869926554 0.0004058834 0.2788778392 0.6281262147
          16           17           18           19           20
0.0246124941 0.6443889709 0.3088615307 0.0369764902 0.0424842835
          21           22           23           24           25
0.0061845776 0.1524780906 0.9941978529 0.1915301359 0.0163126738
          26           27           28
0.3088615307 0.9937780455 0.9733825060
```

Making prediction for new cases

Predict the probability of being hired for a male applicant with 4 years of higher education and 0 years of experience.

$$\ln \frac{\hat{\pi}}{1-\hat{\pi}} = -14.25 + 1.15(4) + .91(0) + 5.6(1) = -4.05$$

$$\hat{\pi} = \frac{e^{-4.05}}{1+e^{-4.05}} = 0.017$$

Predict the probability of being hired for a male applicant with 6 years of higher education and 6 years of experience.

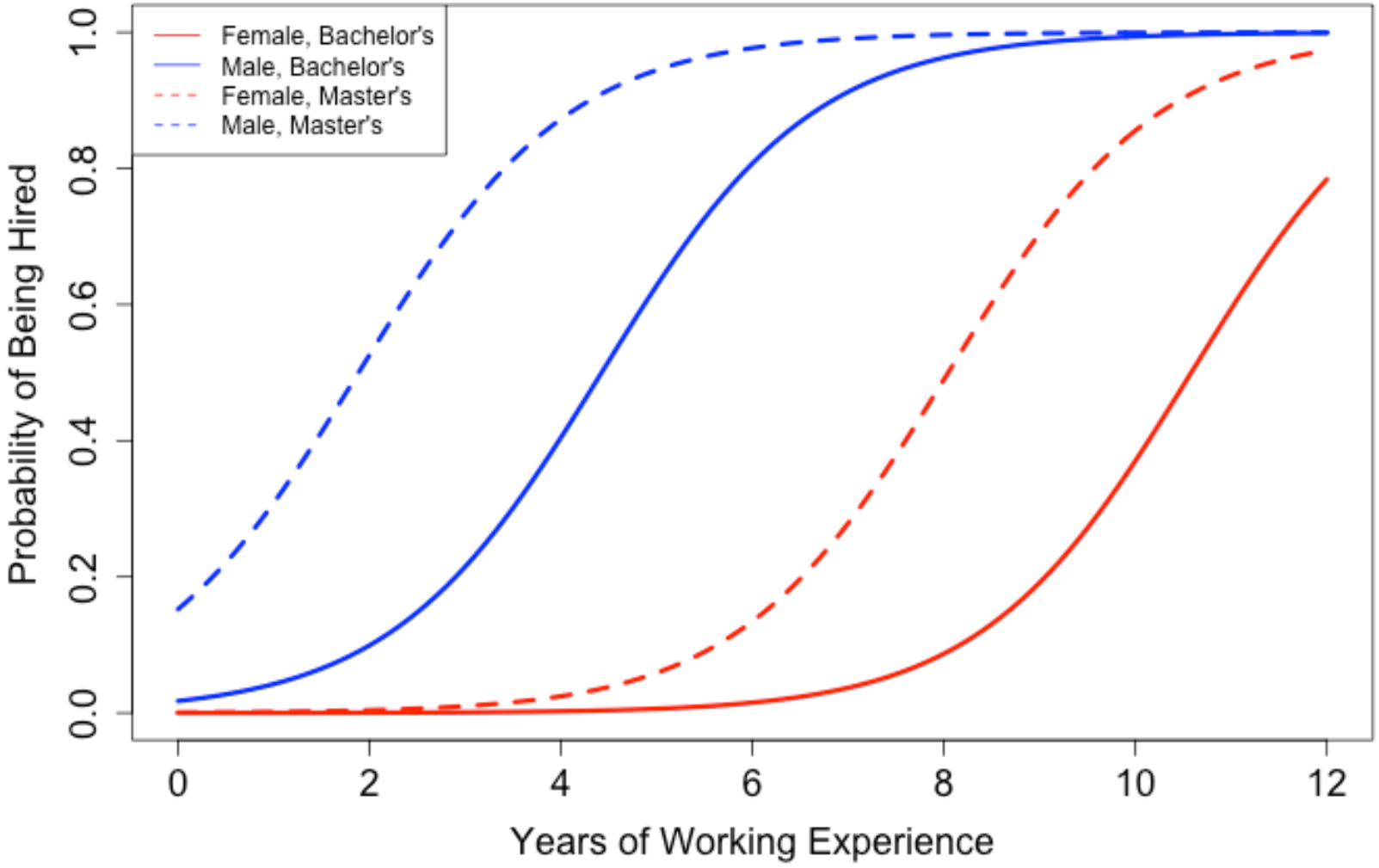
$$\ln \frac{\hat{\pi}}{1-\hat{\pi}} = -14.25 + 1.15(6) + .91(6) + 5.6(1) = 3.71$$

$$\hat{\pi} = \frac{e^{3.71}}{1+e^{3.71}} = 0.976$$

Use R to predict the probability of success for several new cases simultaneously:

```
> my.newdata <- data.frame(EDUC=c(4, 6, 6), EXP=c(0, 6, 6), GENDER=c(1, 1, 0))
> my.newdata
  EDUC EXP GENDER
1    4   0      1
2    6   6      1
3    6   6      0
> predict(m, newdata=my.newdata, type="response")
         1          2          3
0.01754895 0.97688294 0.13471665
```

The Fitted Model Curves



R codes for the previous graph

```
png("hire_graph.png", height=500, width=700)
par(mar=c(6.1, 5.1, 4.1, 2.1))

summary(hire$EXP)
EXP.plot <- seq(0, 12, by=.1)

# GENDER == 0 & EDUC == 4 (Female, Bachelor's degree)

pi <- predict(m, newdata=data.frame(EDUC=4, EXP=EXP.plot, GENDER=0), type="response")
plot(EXP.plot, pi, xlim=c(0, 12), ylim=c(0, 1),
     xlab="Years of Working Experience", ylab="Probability of Being Hired",
     type='l', col='red', lty="solid", lwd=3, cex.axis=1.5, cex.lab=1.5)

# GENDER == 1 & EDUC == 4 (Male, Bachelor's degree)

pi <- predict(m, newdata=data.frame(EDUC=4, EXP=EXP.plot, GENDER=1), type="response")
lines(EXP.plot, pi, col='blue', lty="solid", lwd=3)

# GENDER == 0 & EDUC == 6 (Female, Master's degree)

pi <- predict(m, newdata=data.frame(EDUC=6, EXP=EXP.plot, GENDER=0), type="response")
lines(EXP.plot, pi, col='red', lty="dashed", lwd=3)

# GENDER == 1 & EDUC == 6 (Male, Master's degree)

pi <- predict(m, newdata=data.frame(EDUC=6, EXP=EXP.plot, GENDER=1), type="response")
lines(EXP.plot, pi, col='blue', lty="dashed", lwd=3)

legend(x="topleft",
       legend=c("Female, Bachelor's", "Male, Bachelor's", "Female, Master's", "Male, Master's"),
       col=c("red", "blue", "red", "blue"), lty=c("solid", "solid", "dashed", "dashed"))

dev.off()
```

Outline

1. The problem
2. The logistic regression model
3. The generalized linear models (GLM)
4. Making prediction with the fitted model
5. Interpreting model coefficients
6. Estimating model coefficients
7. Inferences
8. Model utility (practical significance) – pseudo R^2
9. Diagnostics
10. From predicted probability to binary classification

R Output

```
> m <- glm(HIRE~EDUC+EXP+GENDER, data=hire, family=binomial)
> summary(m)

Call:
glm(formula = HIRE ~ EDUC + EXP + GENDER, family = binomial,
    data = hire)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4380  -0.4573  -0.1009   0.1294   2.1804

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.2483     6.0805  -2.343   0.0191 *
EDUC          1.1549     0.6023   1.917   0.0552 .
EXP           0.9098     0.4293   2.119   0.0341 *
GENDER        5.6037     2.6028   2.153   0.0313 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35.165  on 27  degrees of freedom
Residual deviance: 14.735  on 24  degrees of freedom
AIC: 22.735
```

The fitted model:

$$\ln \frac{\pi}{1-\pi} = -14.25 + 1.15(\text{EDU}) + .91(\text{EXP}) + 5.6(\text{GENDERmale})$$

where π is the probability that a job applicant gets hired.

Interpreting model coefficients

Interpretation 1. The slopes b_i: For every 1-unit increase in X_i, the log-odds increases by b_i, holding all other predictors constant.

Examples:

$$\ln \frac{\pi}{1-\pi} = -14.25 + 1.15(\text{EDU}) + .91(\text{EXP}) + 5.6(\text{GENDERmale})$$

$b_1 = 1.15$: Holding years of experience and gender constant, the predicted log-odds that the candidate gets hired increases by 1.15 with every 1 more year of higher education.

$b_3 = 5.6$: Holding years of higher education and years of experience constant, the predicted log-odds that a male candidate gets hired is higher than the log-odds for a female candidate by 5.6.

Interpreting model coefficients

Interpretation 2. e^{b_i}: For every 1-unit increase in X_i, the odds increases by a factor of e^{b_i}, holding all other predictors constant.

Derivation

- Candidate 1: $X_1 = x_1, X_2 = x_2, X_3 = x_3$, predicted probability $\hat{\pi}$.
- Candidate 2: $X_1 = x_1 + 1, X_2 = x_2, X_3 = x_3$, predicted probability $\hat{\pi}^*$.

From interpretation of OLS slopes, we know

$$b_1 = \ln \frac{\hat{\pi}^*}{1 - \hat{\pi}^*} - \ln \frac{\hat{\pi}}{1 - \hat{\pi}} = \ln \frac{\frac{\hat{\pi}^*}{1-\hat{\pi}^*}}{\frac{\hat{\pi}}{1-\hat{\pi}}}$$

Take an exponential on both sides

$$e^{b_1} = \frac{\frac{\hat{\pi}^*}{1-\hat{\pi}^*}}{\frac{\hat{\pi}}{1-\hat{\pi}}} = \text{Odds Ratio}$$

or $\quad$ Odds of Candidate 2 $=$ Odds of Candidate 1 $* \, e^{b_1}$

Interpreting model coefficients

Examples:

$$\ln \frac{\pi}{1-\pi} = -14.25 + 1.15(\text{EDU}) + .91(\text{EXP}) + 5.6(\text{GENDERmale})$$

$e^{b_1} = e^{1.15} = 3.16$: Holding years of experience and gender constant, the predicted odds that the candidate gets hired increases by a factor of 3.16 with every 1 more year of higher education.

$e^{b_3} = e^{5.6} = 270$: Holding years of higher education and years of experience constant, the predicted odds that a male candidate gets hired is 270 times higher than a female candidate.

Interpreting model coefficients

Interpretation 3. $e^{b_i} - 1$ if $b_i > 0$: For every 1-unit increase in x_i, the odds increases by $(e^{b_i} - 1) * 100\%$, holding all other predictors constant.

or...

$1 - e^{b_i}$ if $b_i < 0$: For every 1-unit increase in x_i, the odds decreases by $(1 - e^{b_i}) * 100\%$, holding all other predictors constant.

Interpreting model coefficients

Examples:

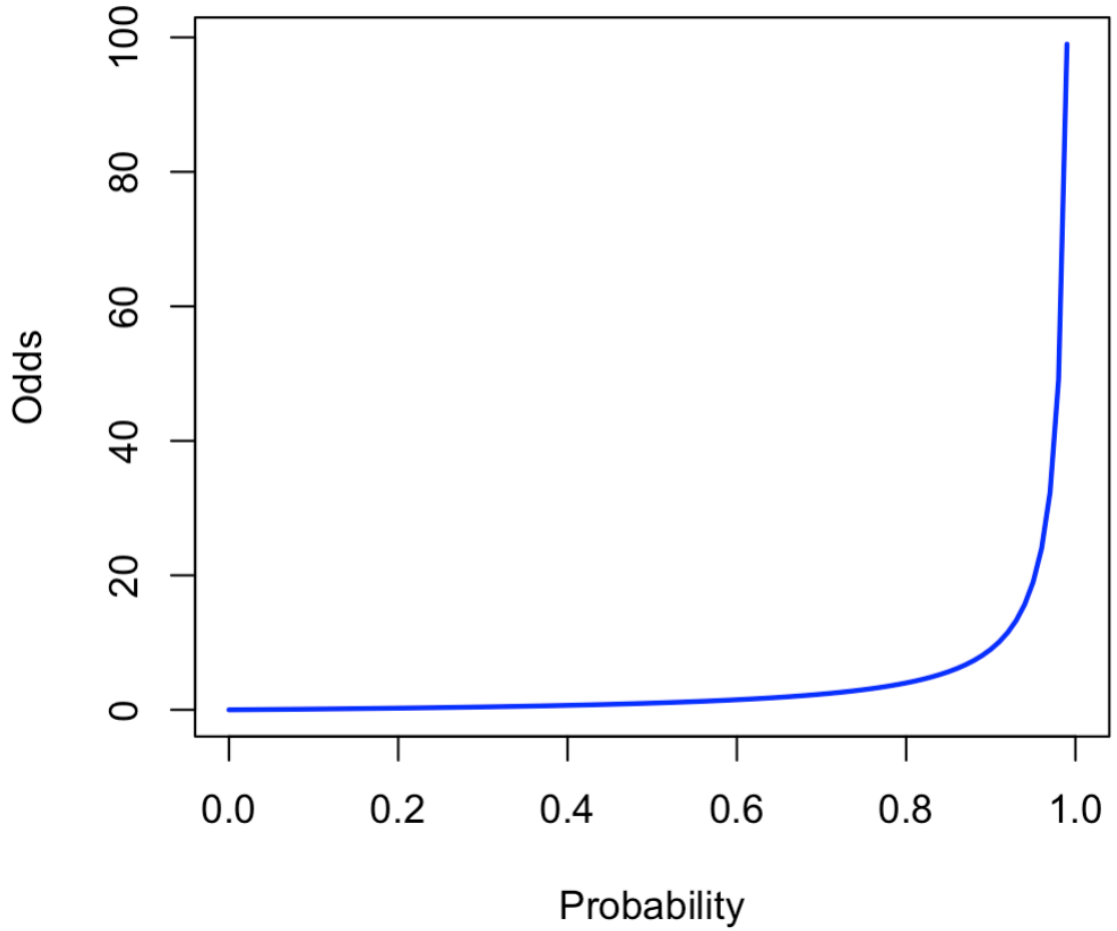
$e^{b_1} = e^{1.15} = 3.16$: Holding years of experience and gender constant, the predicted odds that the candidate gets hired to increase by 216% with every 1 more year of higher education.

We don't have a negative slope in the example data, but let's assume there is a fourth predictor, **length of non-working period (gap) in years**, and $b_4 = -1.5$.

$e^{b_4} = e^{-1.5} = .22$: Holding years of higher education, years of experience, and gender constant, the predicted odds that the candidate gets hired decreases by 78% with every 1 more year of gap.

Keep in mind: The odds is on a different scale than probabilities.

e.g, when probability increases by 1.5 times, how many times does odds increase?



Outline

1. The problem
2. The logistic regression model
3. The generalized linear models (GLM)
4. Making prediction with the fitted model
5. Interpreting model coefficients
6. Estimating model coefficients
7. Inferences
8. Model utility (practical significance) – pseudo R^2
9. Diagnostics
10. From predicted probability to binary classification

Maximum likelihood estimation of logistic regression

Maximize

$$L = \prod_{i=1}^{n} \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i}$$

where

$$\pi_i = \frac{\exp(\mathbf{X_i b})}{1 + \exp(\mathbf{X_i b})}$$

Example

The first five candidates' observed hiring status: 0, 0, 1, 1, 0.

$$L = (1 - \pi_1)(1 - \pi_2)\pi_3\pi_4(1 - \pi_5)$$

Maximum likelihood estimation

$$L = \prod_{i=1}^{n} \pi_i^{Y_i} (1 - \pi_i)^{1 - Y_i}$$

To simplify the optimization problem, we take a natural log of the above likelihood function:

$$\begin{aligned}
\ln L &= \sum_{i=1}^{n} [Y_i \ln \pi_i + (1 - Y_i) \ln(1 - \pi_i)] \\
&= \sum_{i=1}^{n} \left[Y_i \ln \left(\frac{\pi_i}{1 - \pi_i} \right) \right] + \sum_{i=1}^{n} \ln(1 - \pi_i) \\
&= \sum_{i=1}^{n} Y_i (\mathbf{X}_i \boldsymbol{\beta}) - \sum_{i=1}^{n} \ln[1 + \exp(\mathbf{X}_i \boldsymbol{\beta})]
\end{aligned}$$

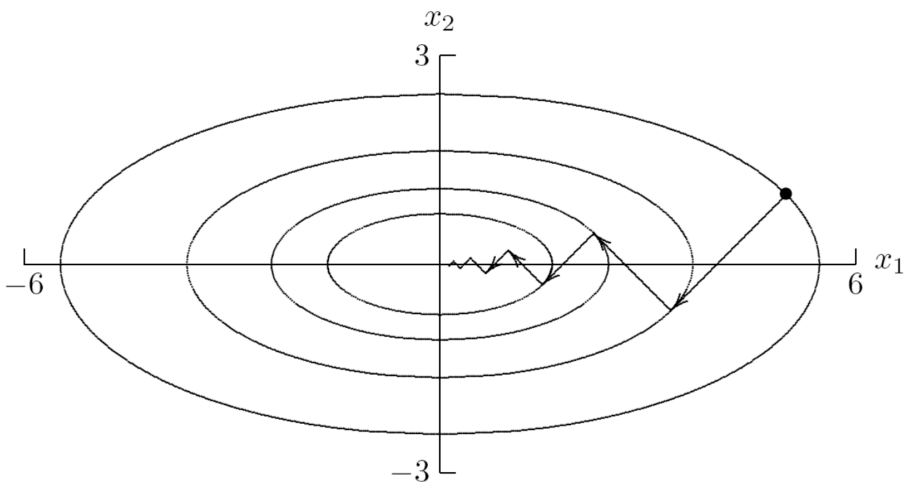
The MLE seeks the vector of coefficients $\hat{\boldsymbol{\beta}}$ that maximizes the log-likelihood.

Maximum likelihood estimation

No closed-form solution exists for the model coefficients. Computer-intensive numerical search procedures are required. Below are two basic ideas of numerical search.

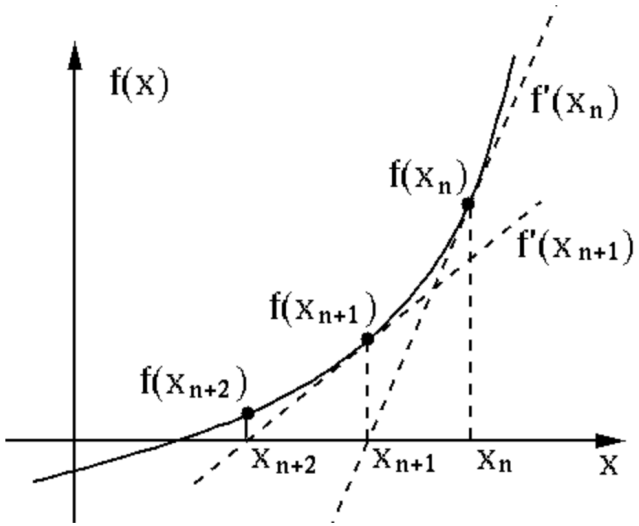
Steepest Descent

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)$$



Newton-Raphson

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{H}_f^{-1}(\boldsymbol{x}_k)\nabla f(\boldsymbol{x}_k)$$



Outline

1. The problem
2. The logistic regression model
3. The generalized linear models (GLM)
4. Making prediction with the fitted model
5. Interpreting model coefficients
6. Estimating model coefficients
7. Inferences
8. Model utility (practical significance) – pseudo R^2
9. Diagnostics
10. From predicted probability to binary classification

The same types of inferences are of interest in logistic regression as for linear regression models:

- Hypothesis tests and confidence intervals of the regression coefficients
- Confidence interval of the estimated mean response given $\mathbf{X}_h$
- Hypothesis test to compare full vs. reduced models

Sampling distribution of estimated model coefficients

The inference procedures of logistic regression rely on **large sample sizes**.

For large samples, ML estimators **b** for logistic regression are:

- Approximately normally distributed
- $E\{b\} = \beta$
- The approximate variances and covariances are functions of the second-order partial derivatives of $\ln L$.

Let **G** denote the matrix of the second-order partial derivatives of $\ln L$ with regard to all regression coefficients $\beta_0, \beta_1, \cdots, \beta_{p-1}$ (a.k.a., the Hessian Matrix):

$$\underset{p\times p}{\mathbf{G}} = [g_{ij}]$$

where

$$g_{00} = \frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial \beta_0^2}$$

$$g_{01} = \frac{\partial^2 \ln L(\boldsymbol{\beta})}{\partial \beta_0 \partial \beta_1}$$

$$etc.$$

When **G** is evaluated at the MLE of $\boldsymbol{\beta}$, the estimated approximate variance of the estimated regression coefficients **b** are:

$$\mathbf{s}^2\{b_k\} = -\frac{1}{g_{kk}|\boldsymbol{\beta} = \mathbf{b}_{MLE}}$$

Confidence interval of model coefficients

Because $(b_k - \beta_k)/s\{b_k\}$ follows an approximate standard normal distribution...

The $1 - \alpha$ confidence interval for β_k is,

$$b_k \pm z_{1-\alpha/2} \cdot s\{b_k\}$$

- b_k and $s\{b_k\}$ can be obtained from the R model summary output.
- $z_{1-\alpha/2}$ can be obtained by R code: `qnorm(p=(1-alpha/2))`.

Interpretations:

- We are $(1-\alpha)100\%$ sure that the true population parameter β_k falls within this interval.
- (For the slope parameters only) With a $(1-\alpha)100\%$ confidence, we estimate that the **log-odds** of success changes by somewhere between LL and UL for each 1-unit increase in X_k while holding all other predictors constant.
- (For the slope parameters only) With a $(1-\alpha)100\%$ confidence, we estimate that the **odds** of success changes by **a factor of** somewhere between e^{LL} and e^{UL} for each 1-unit increase in X_k while holding all other predictors constant.

Hypothesis testing of model coefficients

1. **Assumptions:** large sample, independent observations
2. **Hypotheses:** $H_0 : \beta_k = 0 \qquad H_1 : \beta_k \neq 0$
3. **Test-statistic:**
$$z = \frac{b_k - 0}{s\{b_k\}}$$
4. **P-value:** The two-tail probability in the sampling distribution (z distribution) — If the true slope $\beta_k = 0$, what is the probability that we get a b_k this far from 0 or even more extreme?
5. **Conclusion:** Reject H_0 if p-value is less than a pre-determined significance level (typically 0.05) and conclude that the true population slope is not 0, in other words, X_k is a significant predictor of $E\{Y\} = \pi$.

Confidence interval of mean response $\hat{\pi}_h$

The confidence interval of mean response $\hat{\pi}_h$ is obtained through two steps:

1. Obtain the confidence interval of the logit $\hat{\pi}'_h = \mathbf{X}_h\mathbf{b}$.

$$s^2\{\hat{\pi}'_h\} = \mathbf{X}_h\mathbf{s}^2\{\mathbf{b}\}\mathbf{X}'_h$$

   So the $1-\alpha$ confidence interval is:

$$\hat{\pi}'_h \pm z_{1-\alpha/2} \cdot s\{\hat{\pi}'_h\}$$

2. Transform the lower limit (LL) and the upper limit (UL) of the logit $\hat{\pi}'_h$ to the scale of $\hat{\pi}_h$ by the following formula:

$$\hat{\pi}_h = \frac{\exp(\hat{\pi}'_h)}{1+\exp(\hat{\pi}'_h)}$$

```
> my.pred <- predict(m, newdata=data.frame(EDUC=6, EXP=3, GENDER=1),
+                    level=.95, type="link", se.fit=T)
>
> z.crit <- qnorm(p=.975)
>
> LL.logit <- my.pred$fit - z.crit * my.pred$se.fit
> UL.logit <- my.pred$fit + z.crit * my.pred$se.fit
>
> LL.pi <- exp(LL.logit) / (1 + exp(LL.logit))
> UL.pi <- exp(UL.logit) / (1 + exp(UL.logit))
>
> LL.pi
        1
0.2680399
>
> UL.pi
        1
0.9540469
>
> # Note the point estimator of pi is not the center of the confidence interval
> predict(m, newdata=data.frame(EDUC=6, EXP=3, GENDER=1), type="response")
        1
0.7338511
```

Interpretation: We are 95% sure that the probability of being hired for a male applicant with 6 years of higher education, 3 years of work experience falls between .268 and .954.

Testing whether several $\beta_k = 0$: Likelihood ratio test

Similar to the "general linear test approach", we first specify the full model and the reduced model, where in the reduced model one or more coefficients are set to 0.

Examples:

- The **full model**: $logit = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- The **reduced model**:

$\beta_2 = 0$	which renders	$logit = \beta_0 + \beta_1 X_1$
$\beta_1 = \beta_2 = 0$	which renders	$logit = \beta_0$

Likelihood ratio test

1. Estimate the full model and the reduced model with the same data using MLE.
2. Take the likelihood L for each model evaluated at $\mathbf{b}_{MLE}$, denoted by $L(F)$ and $L(R)$, respectively.
3. Calculate the test statistic

$$G^2 = -2\ln\frac{L(R)}{L(F)} = -2[\ln L(R) - \ln L(F)] = \text{Deviance}(R) - \text{Deviance}(F)$$

Note: $L(R)$ can never be larger than $L(F)$. A large ratio $L(R)/L(F)$ (close to 1), hence a small G^2, indicates that the reduced and full model fit the data equally well. A small ratio $L(R)/L(F)$ (close to 0), hence a large G^2, indicates that the full model fits the data significantly better than the reduced model.

Likelihood ratio test

1. **Assumptions:** large sample, independent observations
2. **Hypotheses:** H_0 : one or more $\beta_k = 0$ $\quad$ H_1 : at least one $\beta_k \neq 0$
3. **Test-statistic:**

$$G^2 = -2 \ln \frac{L(R)}{L(F)} = -2[\ln L(R) - \ln L(F)] = \text{Deviance}(R) - \text{Deviance}(F)$$

4. **P-value:** The right-tail probability in the sampling distribution of G^2 under H_0: χ^2 distribution with a d.f. of the number of parameters that are set to 0 in the reduced model.
5. **Conclusion:** Reject H_0 if p-value is less than a pre-determined significance level (typically 0.05) and conclude that the full model fits the data significantly better than the reduced model; otherwise conclude that the two models fit the data equally well.

Test whether GENDER should be kept in the model

1. **Assumptions:** large sample, independent observations
2. **Hypotheses:** $H_0 : \beta_3 = 0 \qquad H_1 : \beta_3 \neq 0$

   The **full model**: $logit(HIRE) = \beta_0 + \beta_1(EDU) + \beta_2(EXP) + \beta_3(GENDER)$
   The **reduced model**: $logit(HIRE) = \beta_0 + \beta_1(EDU) + \beta_2(EXP)$
3. **Test-statistic:**

$$G^2 = \text{Deviance}(R) - \text{Deviance}(F) = 26.056 - 14.735 = 11.321$$

4. **P-value:** `1 - pchisq(q=11.321, df=25-24)` = 0.000766
5. **Conclusion:** Because the p-value is less than the pre-determined significance level $\alpha = 0.05$, reject H_0 and conclude that the full model fits the data significantly better than the reduced model. GENDER contributes significantly to the model on top of EDU and EXP. There is evidence for gender discrimination after accounting for education and work experience.

The Full Model:

```
> m.F <- glm(HIRE~EDUC+EXP+GENDER, data=hire, family=binomial)
> summary(m.F)

Call:
glm(formula = HIRE ~ EDUC + EXP + GENDER, family = binomial,
    data = hire)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4380  -0.4573  -0.1009   0.1294   2.1804

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.2483     6.0805  -2.343   0.0191 *
EDUC          1.1549     0.6023   1.917   0.0552 .
EXP           0.9098     0.4293   2.119   0.0341 *
GENDER        5.6037     2.6028   2.153   0.0313 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35.165  on 27  degrees of freedom
Residual deviance: 14.735  on 24  degrees of freedom
AIC: 22.735

Number of Fisher Scoring iterations: 7
```

The Reduced Model:

```
> m.R <- glm(HIRE~EDUC+EXP, data=hire, family=binomial)
> summary(m.R)

Call:
glm(formula = HIRE ~ EDUC + EXP, family = binomial, data = hire)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5031  -0.6810  -0.3902   0.5550   1.9886

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.4419     2.3361  -2.330   0.0198 *
EDUC          0.5404     0.3287   1.644   0.0002
EXP           0.3717     0.1670   2.226   0.0260 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35.165  on 27  degrees of freedom
Residual deviance: 26.056  on 25  degrees of freedom
AIC: 32.056

Number of Fisher Scoring iterations: 5
```

Likelihood ratio test of global model utility (fitted vs. null model)

1. **Assumptions:** large sample, independent observations
2. **Hypotheses:** H_0 : All slopes $\beta_k = 0$ $\quad$ H_1 : at least one slope $\beta_k \neq 0$

   The **full model**: $logit = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$

   The **reduced model**: $logit = \beta_0$ (the null model)
3. **Test-statistic:**
$$G^2 = -2 \ln \frac{L(R)}{L(F)} = -2[\ln L(Null) - \ln L(Fitted)]$$
$$= \text{Deviance}(Null) - \text{Deviance}(Residual)$$
4. **P-value:** The right-tail probability in the reference distribution: χ^2 with the df being the difference between the dfs of the two deviances.
5. **Conclusion:** Reject H_0 if p-value is less than a pre-determined significance level (typically 0.05) and conclude that the full model fits the data significantly better than the reduced model. **At least one of the predictors is useful for predicting the binary outcome.**

Outline

1. The problem
2. The logistic regression model
3. The generalized linear models (GLM)
4. Making prediction with the fitted model
5. Interpreting model coefficients
6. Estimating model coefficients
7. Inferences
8. Model utility (practical significance) – pseudo R^2
9. Diagnostics
10. From predicted probability to binary classification

The old R^2 is not applicable to logistic regression because there is no SSE herein. We introduce two "Pseudo R^2s".

McFadden's Pseudo R^2

$$R^2_{MF} = 1 - \frac{-2\ln(L_{\text{fitted}})}{-2\ln(L_{\text{null}})} = 1 - \frac{\text{Residual deviance}}{\text{Null deviance}} = 1 - \frac{14.735}{35.165} = 0.58$$

where L_{fitted} is the likelihood of the fittedmodel and L_{null} is the likelihood of the null model where no predictors are involved.

McFadden's Pseudo R^2 ranges between 0 and 1, with higher value indicating better model fit.

Interpretation of McFadden's Pseudo R^2

You cannot use the old template for interpreting the regular R^2 (i.e., % variation in y explained by the model) to interpret McFadden's R-squared, because the latter is produced through a different mechanism. There is no intuitive interpretation of McFadden's Pseudo R^2 using daily language.

On page 307 of McFadden's book chapter "Quantitative Methods for Analysing Travel Behavior of Individuals: Some Recent Developments" (1979), in the footnotes, McFadden wrote:

"Those unfamiliar with the ρ^2 index should be forewarned that its values tend to be considerable lower than those of the R^2 index and should not be judged by the standards for a 'good fit' in ordinary regression analysis. For example, values of 0.2 to 0.4 for rho^2 represent an excellent fit."

Interpretation of McFadden's Pseudo R^2

On page 124 of McFadden's book "Urban Travel Demand: A Behavioral Analysis" (1975, Chapter 5), he presented the empirical curve of the relationship between McFadden's Pseudo R^2 (ρ^2) and the regular R^2:

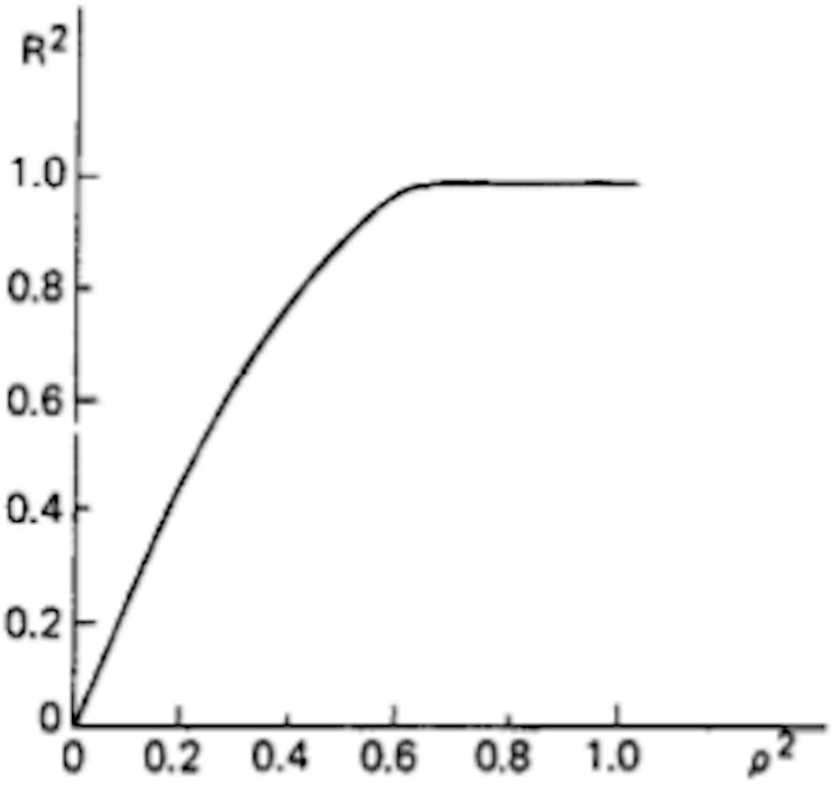


Fig. 5.5.

Tjur (2009) Pseudo R^2

1. Separate data into two groups: the group with observed "success" outcomes, and the group with observed "failure" outcomes.
2. For each group, calculate the mean of the predicted probabilities of success.
3. Take the absolute value of the difference between those two means.

```
> # Pseudo R-square (Tjur, 2009)
>
> sel <- hire$HIRE == 1
>
> abs(mean(predict(m, type="response")[sel]) -
mean(predict(m, type="response")[!sel]))
[1] 0.6210054
```

Outline

1. The problem
2. The logistic regression model
3. The generalized linear models (GLM)
4. Making prediction with the fitted model
5. Interpreting model coefficients
6. Estimating model coefficients
7. Inferences
8. Model utility (practical significance) – pseudo R^2
9. Diagnostics
10. From predicted probability to binary classification

GLM diagnostics

Most of the diagnostics of multiple regression extend to GLMs. These extensions typically take advantage of the computation of the MLE for GLMs by the iteratively reweighted least square (IRLS), which in effect approximates the true log-likelihood by a WLS problem. At the convergence of the IRLS algorithm, diagnostics are formed as if the WLS problem were the problem of interest, and so the exact diagnostics for the WLS fit are approximate diagnostics for the original GLM.

The following functions, some in standard R and some in the `car` package, have methods for GLMs: `rstudent`, `hatvalues`, `cooks.distance`, `dfbeta`, `dfbetas`, `outlierTest`, `marginalModelPlots`, `avPlots`, `residualPlots`, `influencePlot`, `influenceIndexPlot`, `crPlots`, and `ceresPlots`.

See more detail for performing diagnostics for GLMs using the `car` package in this chapter: `www.sagepub.com/upm-data/38503_Chapter6.pdf`.

Logistic regression residuals

Response/raw residuals

$$e_i = Y_i - \hat{\pi}_i$$

Pearson's residuals: e_i divided by estimated standard error of Y_i

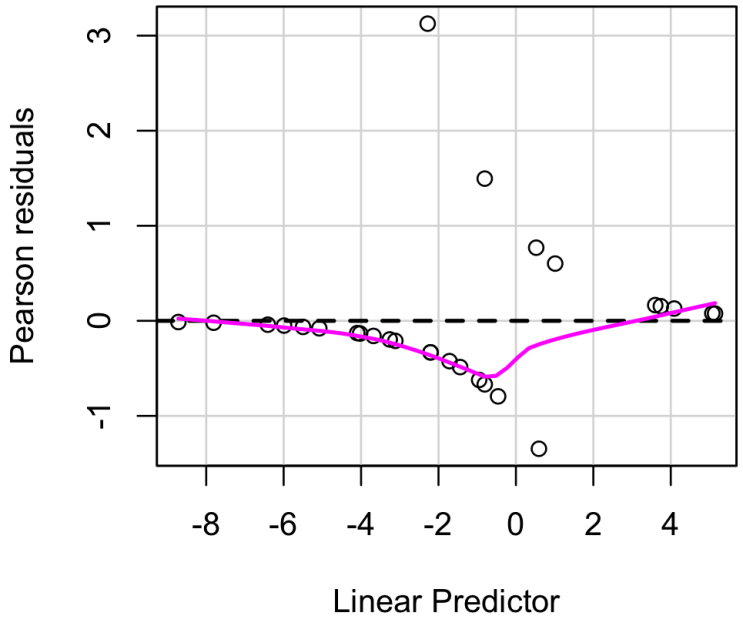
$$r_{Pi} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}$$

Studentized Pearson's residuals

$$r_{SPi} = \frac{r_{Pi}}{\sqrt{1 - h_{ii}}} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)(1 - h_{ii})}}$$

Residual vs. fitted plot

```
car::residualPlots(m, terms= ~1)
```



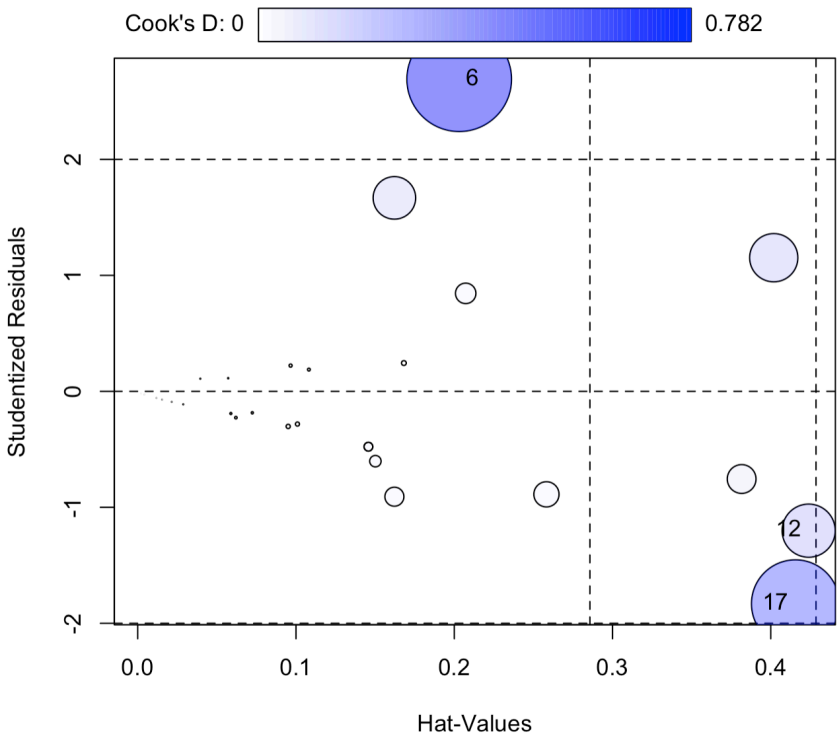
If the logistic regression model is correct, then $E\{Y_i\} = \pi$, and the residuals follow asymptotically that: $E\{e_i\} = E\{Y_i - \hat{\pi}_i\} = 0$. This suggests that if the model is correct, a LOWESS smooth curve of the plot of the residuals against the estimated linear predictor should result approximately in a horizontal line.

Cook's distance and influence plot

Cook's distance for logistic regression measures the standardized change in the logit when the ith case is deleted.

$$D_i = r_{SPi}^2 \cdot \frac{h_{ii}}{1 - h_{ii}} \cdot \frac{1}{p}$$

```
car::influencePlot(m)
```



The largest Cook's D value, .782, corresponds to a 45th percentile on the reference distribution $F(df1 = p, df2 = n - p)$.

```
> p <- 4
> n <- nrow(hire)
> pf(q = .782, df1 = p, df2 = n - p)
[1] 0.4524475
```

DFbetas

```
> data.frame(hire, round(dfbetas(m), 2))
   HIRE EDUC EXP GENDER X.Intercept. EDUC.1 EXP.1 GENDER.1
1     0    6   2      0        -0.01   0.01  0.01     0.02
2     0    4   0      1        -0.05   0.05  0.05     0.04
3     1    6   6      1        -0.08   0.06  0.09     0.08
4     1    6   3      1        -0.33   0.29  0.34     0.44
5     0    4   1      0         0.00   0.00  0.00     0.00
6     1    8   3      0         0.68   0.06 -1.02    -1.35
7     0    4   2      1        -0.18   0.20  0.12     0.07
8     0    4   4      0        -0.01   0.01  0.01     0.01
9     0    6   1      0        -0.01   0.01  0.01     0.01
10    1    8  10      0        -0.07   0.07  0.07     0.05
11    0    4   2      1        -0.18   0.20  0.12     0.07
12    0    8   5      0         0.23  -0.67 -0.07     0.36
13    0    4   2      0         0.00   0.00  0.00     0.00
14    0    6   7      0        -0.10   0.03 -0.09     0.24
15    1    4   5      1        -0.22  -0.14  0.70     0.75
16    0    6   4      0        -0.06   0.04  0.05     0.07
17    0    8   0      1         1.01  -1.47 -0.17    -0.80
18    1    6   1      1         0.17  -0.06 -0.30     0.08
```

Outline

1. The problem
2. The logistic regression model
3. The generalized linear models (GLM)
4. Making prediction with the fitted model
5. Interpreting model coefficients
6. Estimating model coefficients
7. Inferences
8. Model utility (practical significance) – pseudo R^2
9. Diagnostics
10. From predicted probability to binary classification

From predicted probability to binary classification

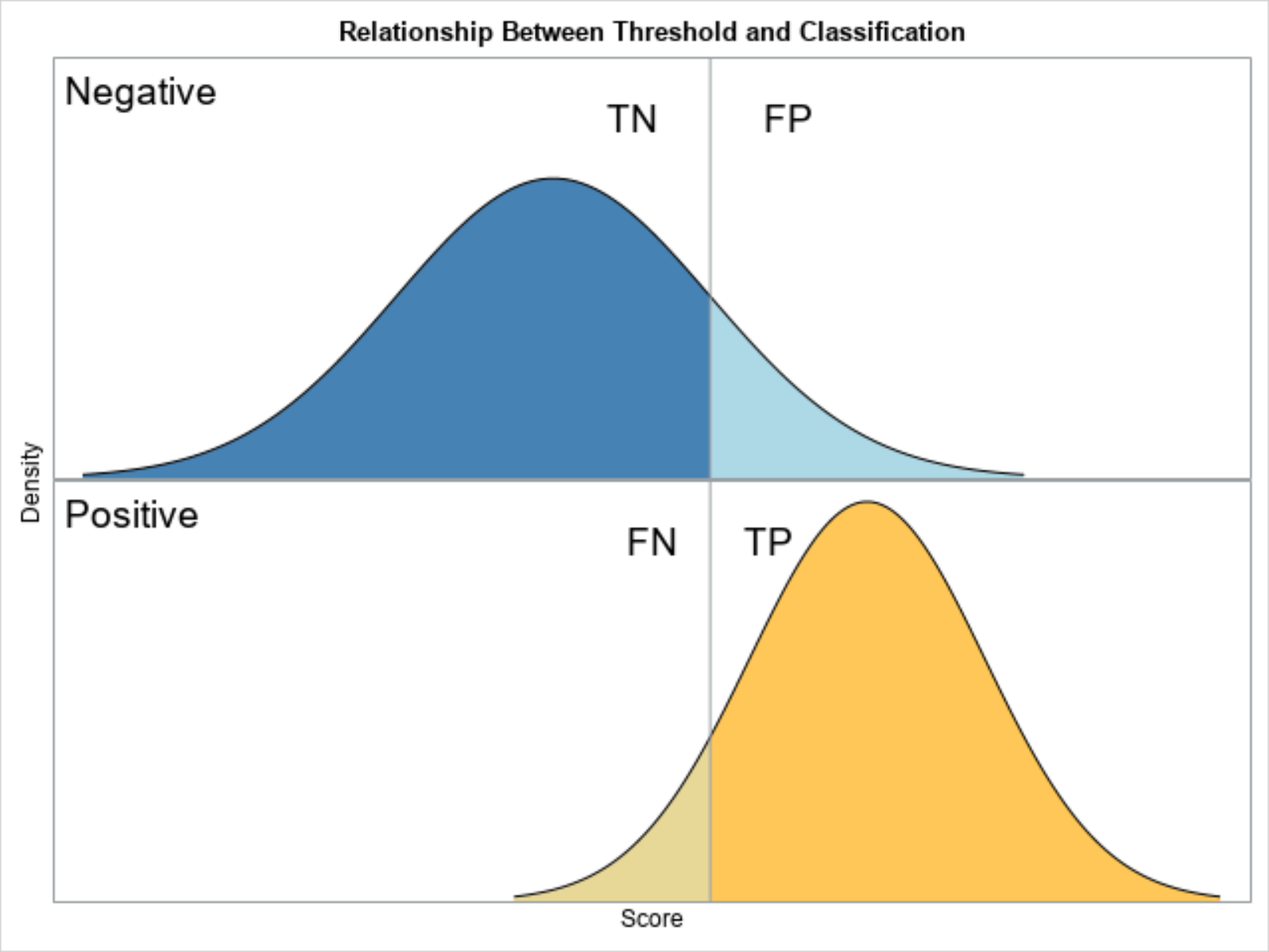
Threshold at $\hat{\pi} = .5$

```
> pi <- predict(m, type="response")
> predicted.class <- rep(0, nrow(hire))
> predicted.class[pi > .5] <- 1
>
> head(data.frame(pi, predicted.class))
            pi predicted.class
1 0.0040730658               0
2 0.0175489472               0
3 0.9768829419               1
4 0.7338510832               1
5 0.0001634423               0
6 0.0928208655               0
>
> observed.class <- hire$HIRE
> confusionMatrix(data=factor(predicted.class),
reference=factor(observed.class), positive="1")
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 18  2
         1  1  7

               Accuracy : 0.8929
                 95% CI : (0.7177, 0.9773)
    No Information Rate : 0.6786
    P-Value [Acc > NIR] : 0.008615

                  Kappa : 0.747

 Mcnemar's Test P-Value : 1.000000

            Sensitivity : 0.7778
            Specificity : 0.9474
         Pos Pred Value : 0.8750
         Neg Pred Value : 0.9000
             Prevalence : 0.3214
         Detection Rate : 0.2500
   Detection Prevalence : 0.2857
      Balanced Accuracy : 0.8626
```

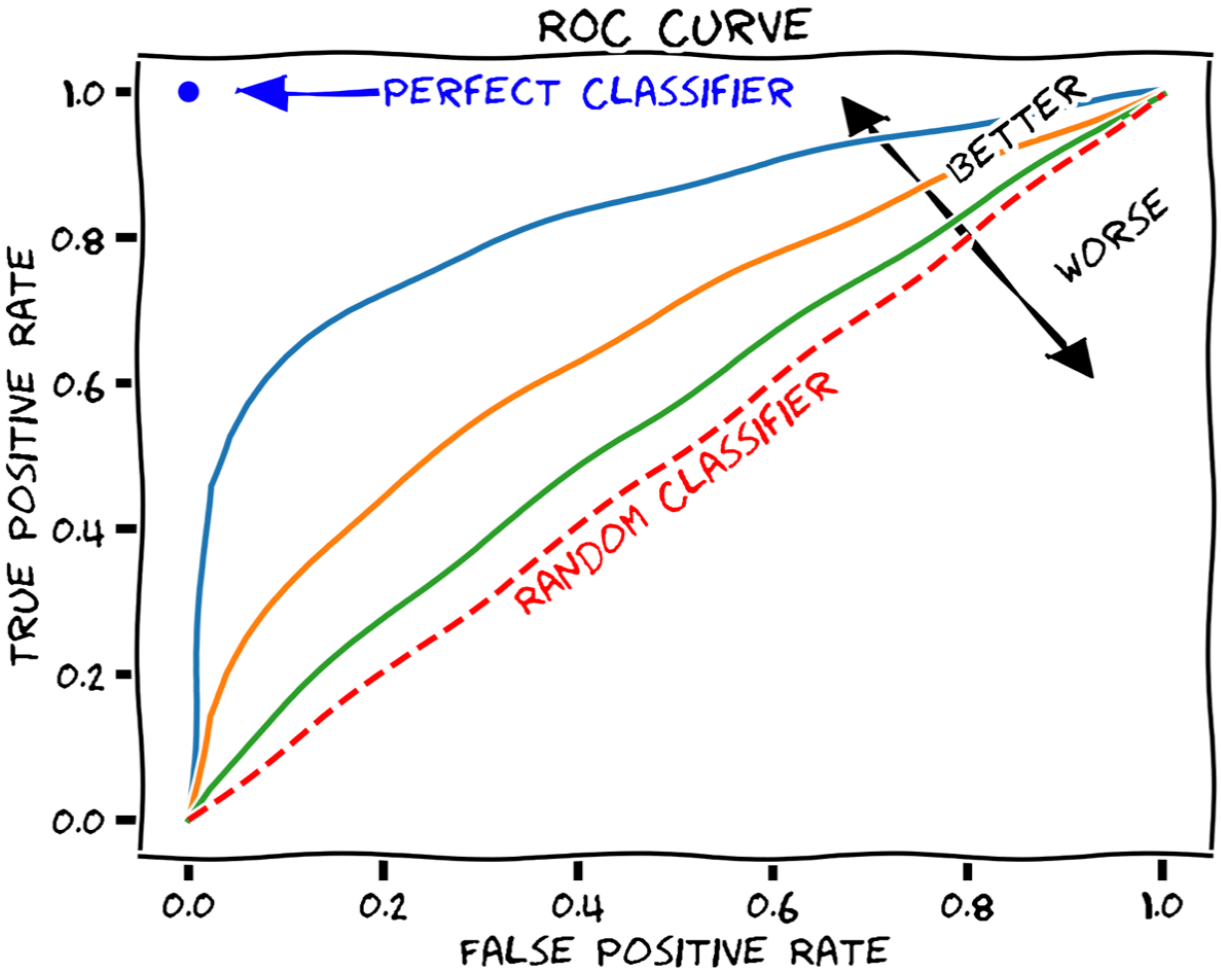
Threshold at $\hat{\pi} = .8$

```
> pi <- predict(m, type="response")
> predicted.class <- rep(0, nrow(hire))
> predicted.class[pi > .8] <- 1
>
> head(data.frame(pi, predicted.class))
            pi predicted.class
1 0.0040730658               0
2 0.0175489472               0
3 0.9768829419               1
4 0.7338510832               0
5 0.0001634423               0
6 0.0928208655               0
>
> observed.class <- hire$HIRE
> confusionMatrix(data=factor(predicted.class),
reference=factor(observed.class), positive="1")
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 19  4
         1  0  5

               Accuracy : 0.8571
                 95% CI : (0.6733, 0.9597)
    No Information Rate : 0.6786
    P-Value [Acc > NIR] : 0.02847

                  Kappa : 0.6291

 Mcnemar's Test P-Value : 0.13361

            Sensitivity : 0.5556
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.8261
             Prevalence : 0.3214
         Detection Rate : 0.1786
   Detection Prevalence : 0.1786
      Balanced Accuracy : 0.7778
```

Binary classification graph



Receiver Operating Characteristic (ROC) curve



Finding an optimal cutpoint based on training data

Options for optimized objective:

- `Youden`: $Se + Sp - 1$
- `MaxProdSpSe`: $Se * Sp$
- `ROC01`:
  distance to the point (0, 1)
- See R documentation of `optimal.cutpoints` for more options and more detail.

```
> library(OptimalCutpoints)
>
> ROC.data <- data.frame(pi, observed.class=hire$HIRE)
>
> out <- optimal.cutpoints(X="pi", status="observed.class", tag.healthy = 0,
+                                 methods="Youden", data=ROC.data)
> out

Call:
optimal.cutpoints.default(X = "pi", status = "observed.class",
    tag.healthy = 0, methods = "Youden", data = ROC.data)

Optimal cutoffs:
  Youden
1 0.3089

Area under the ROC curve (AUC):  0.933 (0.832, 1.034)
>
> tmp <- out$Youden$Global$measures.acc
> Youden.results <- data.frame(tmp$cutoffs, tmp$Se, tmp$Sp, tmp$Se + tmp$Sp - 1)
> colnames(Youden.results) <- c("cutoffs", "Sensitivity", "Specificity", "Youden")
> Youden.results
        cutoffs Sensitivity Specificity     Youden
1  0.0001634423   1.0000000  0.00000000 0.00000000
2  0.0004058834   1.0000000  0.05263158 0.05263158
3  0.0016437556   1.0000000  0.10526316 0.10526316
4  0.0024990525   1.0000000  0.15789474 0.15789474
5  0.0040730658   1.0000000  0.21052632 0.21052632
6  0.0061845776   1.0000000  0.26315789 0.26315789
7  0.0163126738   1.0000000  0.31578947 0.31578947
8  0.0175489472   1.0000000  0.36842105 0.36842105
9  0.0246124941   1.0000000  0.42105263 0.42105263
10 0.0369764902   1.0000000  0.47368421 0.47368421
11 0.0424842835   1.0000000  0.52631579 0.52631579
12 0.0928208655   1.0000000  0.57894737 0.57894737
13 0.0992702542   0.8888889  0.57894737 0.46783626
14 0.1524780906   0.8888889  0.68421053 0.57309942
15 0.1915301359   0.8888889  0.73684211 0.62573099
16 0.2788778392   0.8888889  0.78947368 0.67836257
17 0.3088615307   0.8888889  0.84210526 0.73099415
18 0.3869926554   0.7777778  0.89473684 0.67251462
19 0.6281262147   0.7777778  0.94736842 0.72514620
20 0.6443889709   0.6666667  0.94736842 0.61403509
21 0.7338510832   0.6666667  1.00000000 0.66666667
22 0.9733825060   0.5555556  1.00000000 0.55555556
23 0.9768829419   0.4444444  1.00000000 0.44444444
24 0.9835245625   0.3333333  1.00000000 0.33333333
25 0.9937780455   0.2222222  1.00000000 0.22222222
26 0.9941978529   0.1111111  1.00000000 0.11111111
```