STAT021 Statistical Methods II

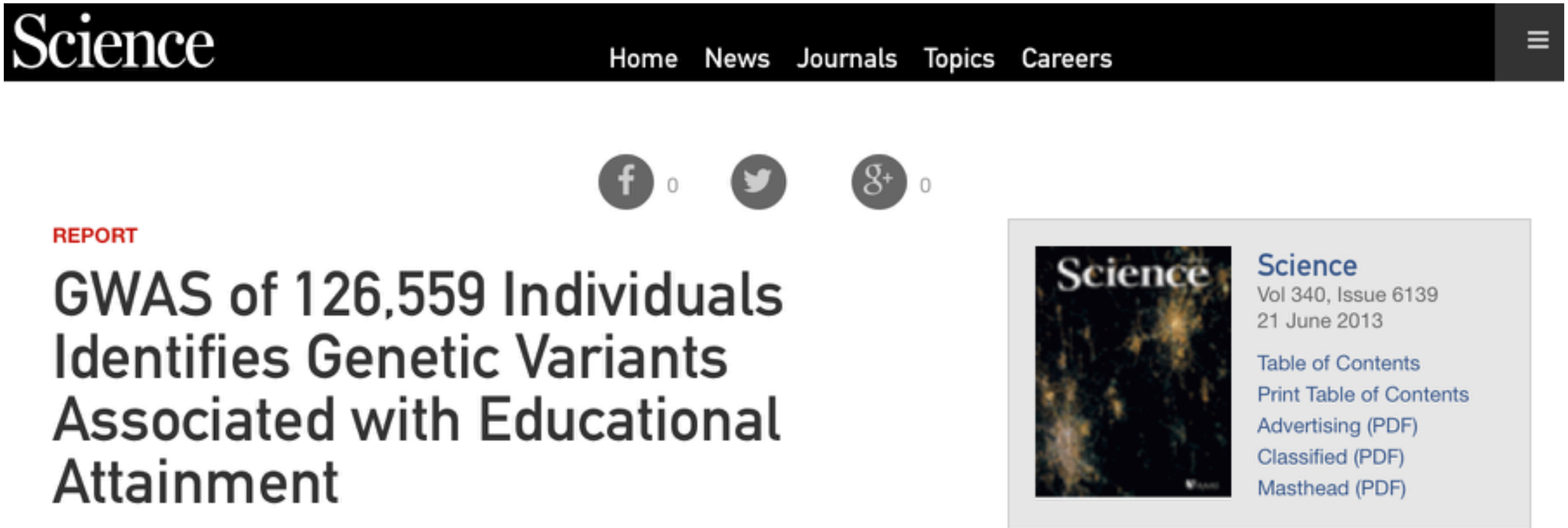
Lecture 22 Logistic Regression

Lu Chen
Swarthmore College
12/4/2018

Outline

- Motivation examples
- Logistic regression
  - Data
  - Bernoulli distribution
  - Definition
  - The logit transformation and the error term
- Probability, odds, log-odds, odds ratio
- Empirical probabilities and estimated probabilities

Motivation examples



- Sample size: 126,559
- Response variables: *EduYears* (Years of schooling) and *College* (college completion)
- Number of SNPs (explanatory variables): ~1,000,000
- Number of tests: ~1,000,000

Motivation examples

SNP	Chr	Discovery stage				Replication stage	
		Beta/OR	*P* value	I^2	P_{het}	Beta/OR	*P* value
				EduYears			
rs9320913	**6**	**0.106**	$\mathbf{4.19\times10^{-9}}$	**18.3**	**0.097**	**0.077**	**0.012**
rs3783006	13	0.096	2.29×10^{-7}	0	0.982	0.056	0.055
rs8049439	16	0.090	7.12×10^{-7}	10.7	0.229	0.065	0.026
rs13188378	5	−0.136	7.49×10^{-7}	0	0.791	0.091	0.914
				College			
rs11584700	**1**	**0.921**	$\mathbf{2.07\times10^{-9}}$	**13.8**	**0.179**	**0.912**	$\mathbf{4.86\times10^{-4}}$
rs4851266	**2**	**1.050**	$\mathbf{2.20\times10^{-9}}$	**23.7**	**0.049**	**1.049**	**0.003**
rs2054125	2	1.468	5.55×10^{-8}	7	0.325	1.098	0.225
rs3227	6	1.043	6.02×10^{-8}	5	0.363	1.010	0.280
rs4073894	7	1.076	4.41×10^{-7}	0	0.765	1.003	0.467
rs12640626	4	1.041	4.94×10^{-7}	10.9	0.234	1.000	0.495

- The response variable of the second analysis is "college completion", which is binary with values $1 = Yes$ and $0 = No$.
- When we are interested in a binary reponse variable, i.e. what genes are related to a person's college completion, we use **Logistic Regression** to model it.

Motivation examples

Go is an abstract strategy board game for two players, in which the aim is to surround more territory than the opponent.

AlphaGo by Google [DeepMind]

> **AlphaGo** uses a Monte Carlo tree search algorithm to find its moves based on knowledge previously "learned" by **machine learning**, specifically by an **artificial** neural network (a **deep learning** method) by extensive training, both from human and computer play.

Motivation examples

Is a teenager's age related to whether he/she sleeps at least 7 hours a night?

- Response variable *Sleep*: whether a teenager sleeps at least 7 hours a night
  $Sleep = 1$ for *Yes* and 0 for *No*.
- Predictor/Explanatory variable *Age*: the age of the teenager, quantitative.

Is a student's GPA related to acceptance to medical schools?

- Response variable *Acceptance*: whether a student is accepted to medical schools
  $Acceptance = 1$ for *Yes* and 0 for *No*.
- Predictor/Explanatory variable *GPA*: quantitative.

Motivation examples

```r
head(TeenSleep,3)
```

```
##   Age Sleep
## 1  16     1
## 2  17     0
## 3  14     1
```

```r
dim(TeenSleep)
```

```
## [1] 446   2
```

```r
head(Med, 3)
```

```
##   Acceptance  GPA
## 1          0 3.62
## 2          1 3.84
## 3          1 3.23
```

```r
dim(Med)
```

```
## [1] 55  2
```

Logistic regression - Data

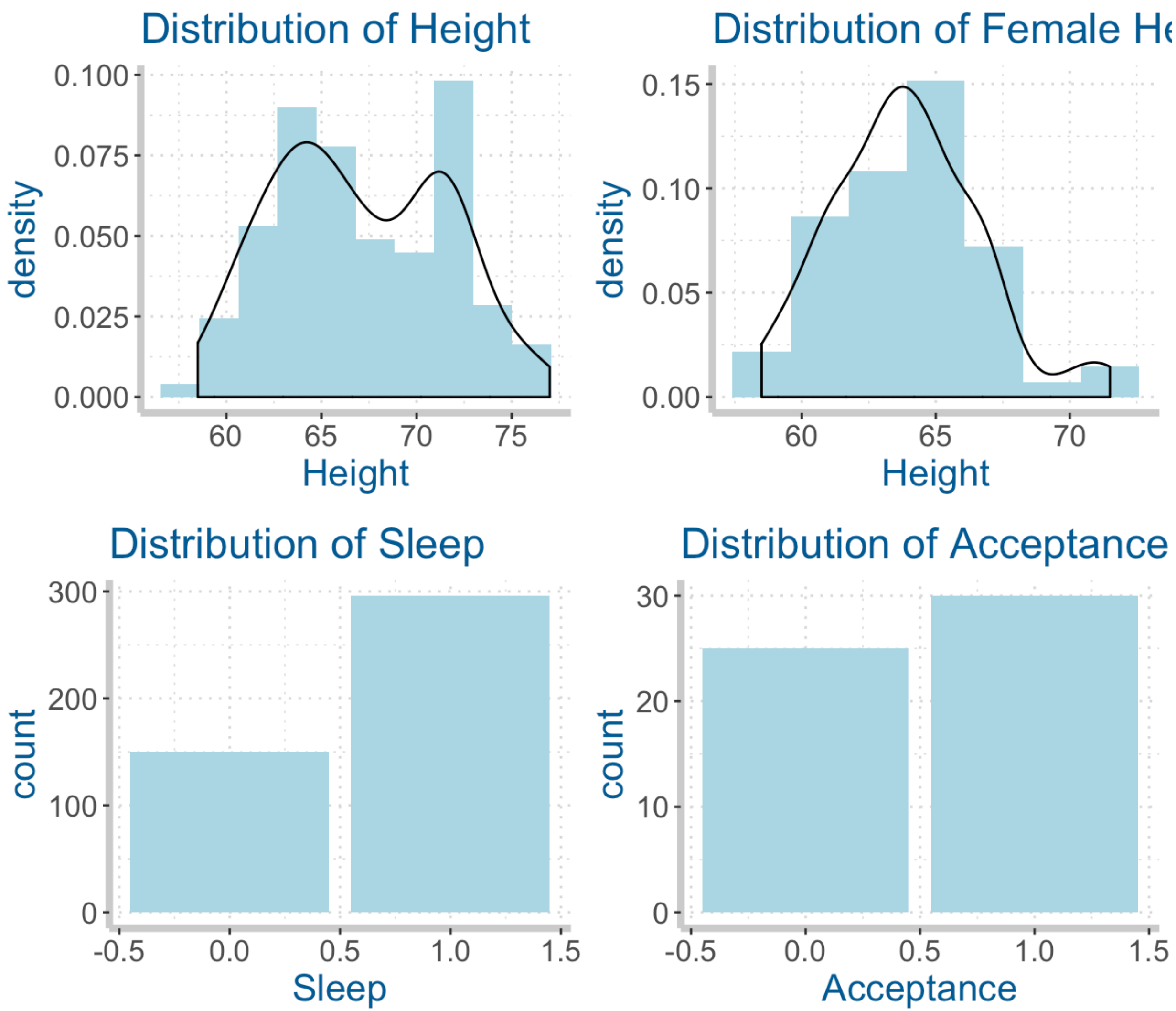
Linear regression

- Response variable Y: quantitative
- Predictors X's: categorical or quantitative
- In real world data, a quantitative Y (or a transformation of it) given a certain X is usually Normally distributed. This is why we assume $Y = \beta_0 + \beta_1 X_1 + \cdots + \epsilon$, where $\epsilon \overset{iid}{\sim} N(0, \sigma)$ in linear regression.

Logistic regression

- Response variable Y: binary, 0 or 1
- Predictors X's: categorical or quantitative
- A binary variable has a completely different data type from a quantitative variable. Therefore, we do not use Normal distribution but a different distribution to describe a binary Y.

Logistic regression - Data



Response variable is quantitative.

Response variable is binary.

Logistic regression - Data

Normal distribution is commonly used to describe quantitative data.

$$Y \sim N(\mu, \sigma)$$

- Parameters: mean μ and SD σ
- In linear regression, μ is a linear function of X's and σ is a constant.

What is the distribution to describe binary data?

Bernoulli distribution.

$$Y \sim Bernoulli(\pi)$$

- Parameter: probability of success π (π is the only parameter)
- $Y = 1$ is called a success; the probability of success is π, $P(Y = 1) = \pi$.
- $Y = 0$ is called a failure; the probability of failure is $1 - \pi$, $P(Y = 0) = 1 - \pi$.
- If $Y \sim Bernoulli(\pi)$, mean of Y is π and SD of Y is $\sqrt{\pi(1 - \pi)}$.

Logistic regression - Bernoulli distribution

Bernoulli distribution $Y \sim Bernoulli(\pi)$.

- Suppose $\pi = 0.2$ and $Y \sim Bernoulli(0.2)$.
- $P(Y = 1) = 0.2$ and $P(Y = 0) = 1 - 0.2 = 0.8$ The probability of getting 1 is 0.2 and the probability of getting 0 is 0.8.
- Suppose there are 10 Y values 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, mean of Y is

$$\mu_Y = \frac{0+0+0+0+0+0+0+0+1+1}{10} = \frac{2}{10} = 0.2 = \pi$$

- π is the probability of success (probability that $Y = 1$). It is also the mean of Y.
- The SD of Y can also be derived using the formula of calculating standard deviation. $\sigma_Y = \sqrt{\pi(1-\pi)} = \sqrt{0.2 \times 0.8} = 0.4$.
- In logistic regression, we will model mean of Y, π, as a function of X's.

Logistic regressoin - Model

The **logistic regression model** for the probability of success π of a binary response variable Y based on predictors $X_1, X_2, \cdots, X_K$ has either of two equivalent forms:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$$

or

$$\pi = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_K X_K}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_K X_K}}$$

Here $\pi = P(Y = 1|X_1, X_2, \cdots, X_K)$ is a **probability** with $0 < \pi < 1$.

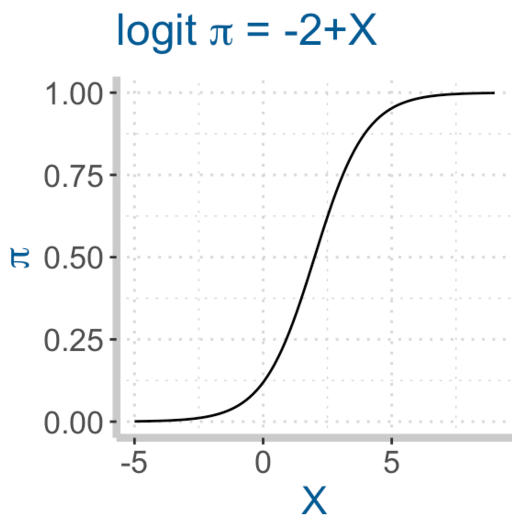
- log is the natural logarithm function with base e.
- The transformation from π to $\log\left(\frac{\pi}{1-\pi}\right)$ is called the **logistic** or **logit** transformation (pronounced "low-JIS-tic" or "LOW-jit").

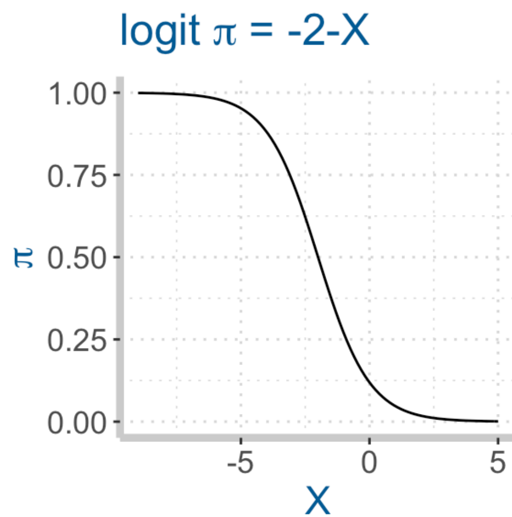
Logistic regressoin - Model

Question 1: Why the logit transformation $\log(\frac{\pi}{1-\pi})$?

$$\text{logit}\,\pi = \log\frac{\pi}{1-\pi} = \beta_0 + \beta_1 X \iff \pi = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

- π is a **monotone non-linear** function of X.
- Slope > 0, π increases as X increases.
- Slope < 0, π decreases as X increases.
- With the transformation, for $-\infty < \beta_0 + \beta_1 X < \infty$, $\pi = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$ is **always between 0 and 1**.





Logistic regressoin - Model

Question 2: Where is the error term?

- There is NO error term in the expression of the logistic regression model.
- π is the only parameter in Bernoulli distribution (knowing it allows us to know both the mean and SD of Y), while in linear regression, we need to estimate both parameters μ and σ in the Normal distribution.
- But no error term does not mean there is no randomness in the data.

For the teenagers' sleeping hours example

- "**State the logistic regression model**": denote *Sleep* as Y and *Age* as X

$$\log \frac{\pi}{1-\pi} = \beta_0 + \beta_1 X, \text{ where } \pi = P(Y = 1 \mid X)$$

- π is the probability that a teenager at age X sleeps at least 7 hours a night.

Logistic regression in R

```r
summary(m1 <- glm(Sleep ~ Age, family="binomial", data=TeenSleep))
```

```
##
## Call:
## glm(formula = Sleep ~ Age, family = "binomial", data = TeenSleep)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6205  -1.4161   0.8443   0.8991   1.0152
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.11864    1.33375   2.338   0.0194 *
## Age         -0.15136    0.08235  -1.838   0.0661 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 569.60  on 445  degrees of freedom
## Residual deviance: 566.19  on 444  degrees of freedom
## AIC: 570.19
```

- $b_0 = 3.12$
- $b_1 = -0.15$

Logistic regression - Estimation

Estimated regression line

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 3.12 - 0.15x \quad \text{or} \quad \hat{\pi} = \frac{e^{3.12-0.15x}}{1+e^{3.12-0.15x}}$$

- $b_1 = -0.15 < 0$, the older the teenagers, the less likely they sleep at least 7 hours a night.
- $x = 14, \hat{\pi} = \hat{P}(Y = 1 \mid X = 14) = 0.73$
  - $\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 3.12 - 0.15 \times 14 = 1.02 \Rightarrow \hat{\pi} = \frac{e^{1.02}}{1+e^{1.02}} = 0.73$
  - About 73% of the teenagers at age 14 sleep at least 7 hours a night.
- $x = 15, \hat{\pi} = \hat{P}(Y = 1 \mid X = 15) = 0.70$
- $x = 16, \hat{\pi} = \hat{P}(Y = 1 \mid X = 16) = 0.67$
- $x = 17, \hat{\pi} = \hat{P}(Y = 1 \mid X = 17) = 0.63$
- $x = 18, \hat{\pi} = \hat{P}(Y = 1 \mid X = 18) = 0.60$

Logistic regression - Estimation

```r
# Estimated log(pi/(1-pi))
logit_pi <- predict(m1)
# Estimated probabilities
est_pi <- exp(logit_pi)/(1+exp(logit_pi))
head(data.frame(TeenSleep, est_pi))
```

```
##   Age Sleep    est_pi
## 1  16     1 0.6674971
## 2  17     0 0.6330972
## 3  14     1 0.7309810
## 4  15     0 0.7001990
## 5  18     1 0.5972855
## 6  14     0 0.7309810
```

- Teenager 1 is younger and sleeps more than teanager 2. The model also estimates that teenager 1 is more likely to sleep at least 7 hours than teenager 2.
- However, teanager 5 is the oldest and predicted to be the leastly likely to sleep at least 7 hours, while in fact, he/she sleeps at least 7 hours a night.
- Teenager 3 and 6 have the same age and thus the same estimated probability, but the observations are different.
- How do we know if the estimations are good or not?

Empirical and estimated probabilites

Age	14	15	16	17	18
$Sleep = 0$	12	35	37	39	27
$Sleep = 1$	34	79	77	65	41
Total	46	114	114	104	68
$\hat{p} = \frac{\#\,Sleep=1}{Total}$	$\frac{34}{46} = 0.739$	$\frac{79}{114} = 0.693$	$\frac{77}{114} = 0.675$	$\frac{65}{104} = 0.625$	$\frac{41}{68} = 0.603$
$\hat{\pi} = \frac{e^{3.12-0.15x}}{1+e^{3.12-0.15x}}$	0.731	0.700	0.667	0.633	0.597

- **Empirical probability $\hat{p}$** is the **observed** $P(Sleep = 1|Age)$ calculated directly from the data.
- **Estimated probability $\hat{\pi}$** is the **estimated** $P(Sleep = 1|Age)$ calculated from the logistic regression model.

Empirical and estimated probabilites

```r
# Empirical probabilities
counts <- table(TeenSleep$Sleep, TeenSleep$Age); counts
```

```
##
##     14 15 16 17 18
##   0 12 35 37 39 27
##   1 34 79 77 65 41
```

```r
prop.table(counts, margin=2)
```

```
##
##            14        15        16        17        18
##   0 0.2608696 0.3070175 0.3245614 0.3750000 0.3970588
##   1 0.7391304 0.6929825 0.6754386 0.6250000 0.6029412
```

```r
# Estimated probabilities
logit_pi <- predict(m1, list(Age=14:18)) # b0+b1x
exp(logit_pi)/(1+exp(logit_pi)) # pi
```

```
##         1         2         3         4         5
## 0.7309810 0.7001990 0.6674971 0.6330972 0.5972855
```

Logistic regression - The slope

Estimated regression line

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 3.12 - 0.15x \quad \text{or} \quad \hat{\pi} = \frac{e^{3.12-0.15x}}{1+e^{3.12-0.15x}}$$

What's the intuitive meaning of the value $b_1 = -0.15$?

- Denote the probability that a teenager at age 14 sleeps at least 7 hours a night as $\hat{\pi}_{14}$ and the probability that a teenager at age 15 sleeps at least 7 hours a night as $\hat{\pi}_{15}$.
- $\log\left(\frac{\hat{\pi}_{14}}{1-\hat{\pi}_{14}}\right) = 3.12 - 0.15 \times 14$and $\log\left(\frac{\hat{\pi}_{15}}{1-\hat{\pi}_{15}}\right) = 3.12 - 0.15 \times 15$
- $b_1 = -0.15 = \log\left(\frac{\hat{\pi}_{15}}{1-\hat{\pi}_{15}}\right) - \log\left(\frac{\hat{\pi}_{14}}{1-\hat{\pi}_{14}}\right)$
- $$b_1 = -0.15 = \log\frac{\hat{\pi}_{15}/(1-\hat{\pi}_{15})}{\hat{\pi}_{14}/(1-\hat{\pi}_{14})} \text{ or } e^{b_1} = e^{-0.15} = \frac{\hat{\pi}_{15}/(1-\hat{\pi}_{15})}{\hat{\pi}_{14}/(1-\hat{\pi}_{14})}$$

Probability, odds and odds-ratio

For any probability π, define

$$\text{Odds} = \frac{\pi}{1-\pi} \text{ and log-odds or logit } \pi = \log \frac{\pi}{1-\pi}$$

The ratio of two odds is defined as odds ratio.

- For example, toss a coin (A), the probability of getting a head is 0.6. $\pi = 0.6$.
- The probability of getting a tail is 0.4. $1 - \pi = 1 - 0.6 = 0.4$.
- The **odds** of getting a head is $Odds_A = \frac{\pi}{1-\pi} = \frac{0.6}{1-0.6} = 1.5$.
- Toss another coin (B), the probability of getting a head is 0.5. Then $Odds_B = \frac{0.5}{1-0.5} = 1$.
- The **log-odds** of getting a head for the two coins are log(1.5) and log(1).
- The **odds ratio** is $\frac{Odds_A}{Odds_B} = \frac{0.6/(1-0.6)}{0.5/(1-0.5)} = \frac{1.5}{1} = 1.5$.

Logistic regression - The slope

Estimated regression line

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 3.12 - 0.15x \quad \text{or} \quad \hat{\pi} = \frac{e^{3.12-0.15x}}{1+e^{3.12-0.15x}}$$

What's the intuitive meaning of the value $b_1 = -0.15$?

$$b_1 = -0.15 = \log\frac{\hat{\pi}_{15}/(1-\hat{\pi}_{15})}{\hat{\pi}_{14}/(1-\hat{\pi}_{14})} \text{ or } e^{b_1} = e^{-0.15} = \frac{\hat{\pi}_{15}/(1-\hat{\pi}_{15})}{\hat{\pi}_{14}/(1-\hat{\pi}_{14})}$$

- $b_1 = -0.15$ is the **difference** in the log-odds that a teenager sleeps at least 7 hours a night between a 15 year old and a 14 year old.
- $e^{b_1} = e^{-0.15}$ is the **ratio** of the odds that a teenager at age 15 sleeps at least 7 hours a night to the odds that a teenager at age 14 sleeps at least 7 hours a night.
- $b_1 > 0$, then $e^{b_1} > 1$; $b_1 < 0$, then $e^{b_1} < 1$; $b_1 = 0$, then $e^{b_1} = 1$.

The medical school acceptance example

```r
head(Med)
```

```
##   Acceptance  GPA
## 1          0 3.62
## 2          1 3.84
## 3          1 3.23
## 4          1 3.69
## 5          1 3.38
## 6          1 3.72
```

Denote *GPA* as X and *Acceptance* as Y. The logistic regression model is

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$$

where $\pi = P(Y = 1|X)$.

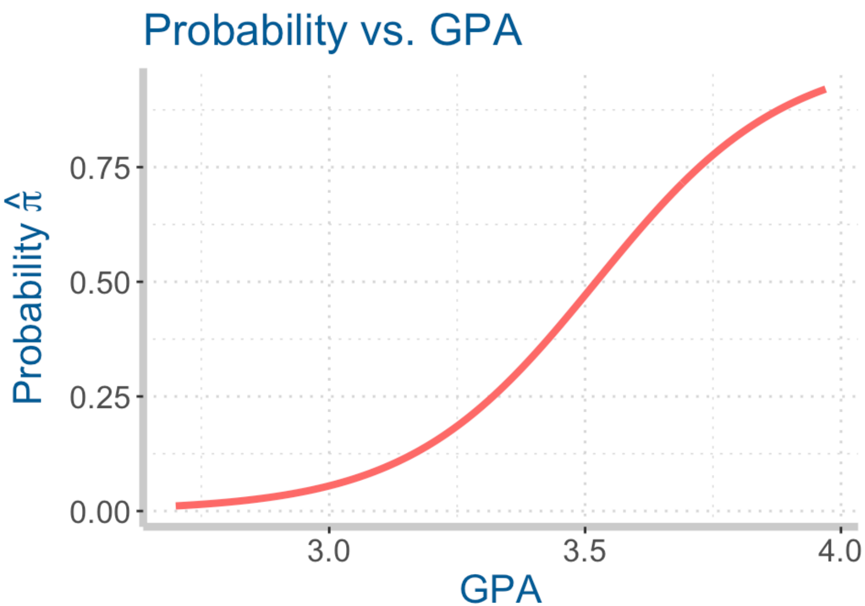
The medical school acceptance example

```r
summary(m2 <- glm(Acceptance ~ GPA, family="binomial", data=Med))
```

```
##
## Call:
## glm(formula = Acceptance ~ GPA, family = "binomial", data = Med)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7805  -0.8522   0.4407   0.7819   2.0967
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -19.207      5.629  -3.412 0.000644 ***
## GPA            5.454      1.579   3.454 0.000553 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 75.791  on 54  degrees of freedom
## Residual deviance: 56.839  on 53  degrees of freedom
## AIC: 60.839
```

- $b_0 = -19.21$
- $b_1 = 5.45$

The medical school acceptance example



$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -19.21 + 5.45x$$

$$\hat{\pi} = \frac{e^{-19.21+5.45x}}{1+e^{-19.21+5.45x}}$$

- Slope $b_1 = 5.45 > 0$, the chance of being accepted by medical schools increases as *GPA* increases.
- As *GPA* gets close to 4.0 or smaller than 3.0, the probability of acceptance approaches 0.93 or 0.
- In the dataset, *GPA* values are from 2.72 to 3.97. Almost all students have different *GPA* values.
- Therefore, it is impossible to calculate the **empirical probabilities** like in the *Sleep* ~ *Age* example.
- But the estimated probabilities can still be obtained from the logistic regression model.

The medical school acceptance example

```r
# Estimated log(pi/(1-pi))
logit_pi <- predict(m2, list(GPA=c(3, 3.5, 3.9))); logit_pi
```

```
##          1          2          3
## -2.8440051 -0.1169222  2.0647442
```

```r
# Estimated pi
est_pi <- exp(logit_pi)/(1+exp(logit_pi)); est_pi
```

```
##          1          2          3
## 0.05499203 0.47080271 0.88742898
```

```r
# Estimated pi for all GPA values in the data set
logit_pi_all <- predict(m2)
est_pi_all <- exp(logit_pi_all)/(1+exp(logit_pi_all))
head(est_pi_all)
```

```
##         1         2         3         4         5         6
## 0.6312488 0.8503685 0.1694476 0.7149136 0.3161716 0.7470602
```

The medical school acceptance example

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -19.21 + 5.45x \text{ or } \hat{\pi} = \frac{e^{-19.21+5.45x}}{1+e^{-19.21+5.45x}}$$

What does the value of of the slope $b_1 = 5.45$ mean?

- As *GPA* increases 1 unit, log-odds of being accepted by medical schools increases 5.45 units.
- $e^{b_1} = e^{5.45} = 233.76$. The odds of being accepted by medical schools is 233.76 times higher for every 1 unit increase in *GPA*.
- 1 unit increase in *GPA* seems too dramatic. Another way to interpret the slope: $e^{b_1 \times 0.1} = e^{0.545} = 1.72$. The odds of being accepted by medical schools is 1.72 times higher for every 0.1 unit increase in *GPA*.
- The **odds ratio e^{b_1}** measures the effect of the predictor X on the response variable Y **multiplicatively** (not additively).

Summary

- **Binary response variable** $Y = 1$ or 0.
- **Bernoulli distribution** for binary data $Y \sim Bernoulli(\pi)$
  - $\pi = P(Y = 1)$; mean of Y is π and SD of Y is $\sqrt{\pi(1-\pi)}$.
- **Logistic regression model**

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K \text{ or } \pi = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_K X_K}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_K X_K}}$$

  where $\pi = P(Y = 1|X_1, X_2, \cdots, X_K)$.
- **Probability** π, **odds** $\frac{\pi}{1-\pi}$, **log-odds** $\log\frac{\pi}{1-\pi}$ and **odds ratio** (ratio of two odds).
- **Empirical probability** (from data) and **estimated probability** (from the logistic regression model).