Lecture 10

Logistic regression: categorical variables

逻辑回归：分类变量

Logistic regression

逻辑回归实现了从基本(基于最小二乘)一般线性模型到中级/高级一般线性模型的过渡

- Logistic regression makes the transition from the basic (least-square-based) *general linear model* to the intermediate/advanced *generalised linear model*

广义线性模型将线性回归模型扩展到非正态分布变量和非线性关系

- The generalised linear model extends linear regression models to variables not normally distributed, and to non-linear relationships

例如，我们可能希望使用回归技术来预测二元响应

- For example, we may want to use regression techniques to predict *binary* responses:
    - we may want to predict the probability that someone is dead or alive, voted or did not vote in the last election etc. as a function of other variables (age, smoking, income etc.)

换句话说，我们仍然希望使用回归，二元概率的结果

- In other words, we still want to use a regression:

$$\text{Probability of binary outcome} = a + b_1X_1 + b_2X_2 \ldots + b_nX_n = a + \Sigma b_iX_i$$

with
a = intercept 截距
b_i = regression coefficients 回归系数
X_i = independent variables (continuous or categorical) 自变量(连续变量或分类变量)

Applications

逻辑回归是一种几乎普遍使用的分类方法，他以概率p预测结果是否发生(二元结果yes/no)

- Logistic regression is a classification method used almost universally
  - it predicts whether an outcome happens or not (binary outcome yes/no) with a probability *p*

他经常用于预测二元结果

- It is frequently applied to predict binary outcomes (yes or no)
  - business: costumer choice (purchasing, being late for bills etc.)
  - medicine, pubic health (will develop a condition etc)
  - insurance (risk of event, credit decisions)
  - etc.

逻辑回归与神经网络和机器学习密切相关

- Logistic regression is closely linked to neural networks and machine learning

Odds and log(odds)

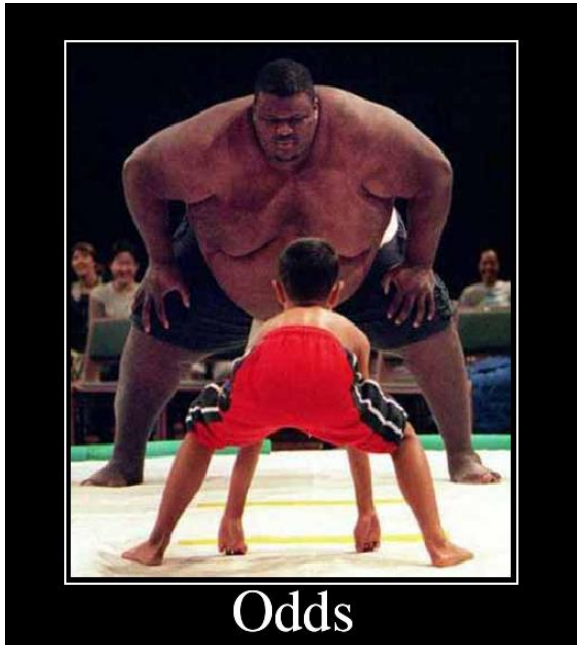
要理解逻辑回归，首先我们需要理解概率和概率比

- To understand logistic regressions, first we need to understand *odds* and *odds ratios*

概率odd和事件的概率并不相同

- Important: odds are not the same as the *probability* of the event!

$$\text{odds of event} = \frac{probability\ of\ event\ occurring}{probability\ of\ event\ not\ occurring}$$



Odds and log(odds)

- Example: what is the *probability* of your birthday falling on a weekday this year?

  - probability of weekday=5/7=0.71 = p

p大于0.5时，odd大于1
log(odd)接近于1

$$\text{Odds of a weekday} = \frac{\text{probability of weekday}}{\text{probability of weekend day}}$$

p小于0.5时，odd小于1
log(odd)是负的

  - odds of weekday = (5/7) / (2/7) = 5/2 = 2.5 = p/(1-p)
  - ln(odds of weekday) = log(2.5) = 0.91 = log(p/(1-p))

非事件的概率

- And the probability of the non-event, i.e. weekend day?
  - probability of weekend day = 2/7=0.29 =1-p
  - odds of weekend day = 2/5 =0.4 =(1-p)/p
  - ln(odds of weekend day)= –0.91 = ln((1-p)/p)

Exercises

Calculate:

- Tossing a fair coin:
  - Probability of heads? 0.5
  - Odds of heads? 1
  - Odds of tails? 1
  - log(odds of heads) 0

- Now throwing a die:
  - Probability of 1? 1/6
  - Odds of 1? 1/5
  - Odds of *not 1*? 5
  - log(odds of 1)? log(0.2)

概率比

Odds ratio

- Now imagine you must choose between betting on coins (bet on 'heads') or dice (bet on '1'); what are the odds of winning in each?
  - odds of heads = 1/1 = 1
  - odds of a 1 = 1/5 =0.2
- So it is easier to win a coin toss; but how much easier?
- We can calculate the odds ratio of success in coins vs. dice
- Odds ratio $= \frac{odds\ of\ heads}{odds\ of\ 1} = \frac{1}{0.2} = 5$
- This means you are 5 times more likely to win by tossing a coin than throwing a die

这意味着通过投硬币比投骰子获胜的可能性高5倍

Summary

概率p介于0和1之间

- probability p is always between 0 and 1

优势和优势比是从0到无穷大

- odds and odds ratio: 0 to +∞

- log(odds) and log(odds ratio): – ∞ to +∞

Odd and probabilities

- If odds = p/(1-p), then:

- p = odds(1-p)
- p = odds – odds*p
- p + odds*p = odds
- p(1 + odds) = odds
- p = odds/(1 + odds)

Therefore

- $p = \frac{1}{1+\frac{1}{odds}}$

(这一页大概就是推导了概率p的计算方法

Break

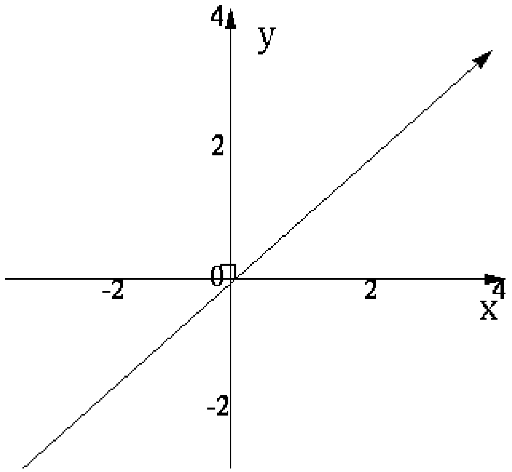
逻辑函数

Logistic function

我们希望使用回归模型从一组预测中计算二元事件的概率

- Back to logistic regression: we want to use a regression model to calculate the probability of binary events (dead/alive, head/tail etc.) from a set of predictors:

$$y = a + b_1X_1 + b_2X_2 \ldots + b_nX_n = a + \Sigma b_iX_i$$

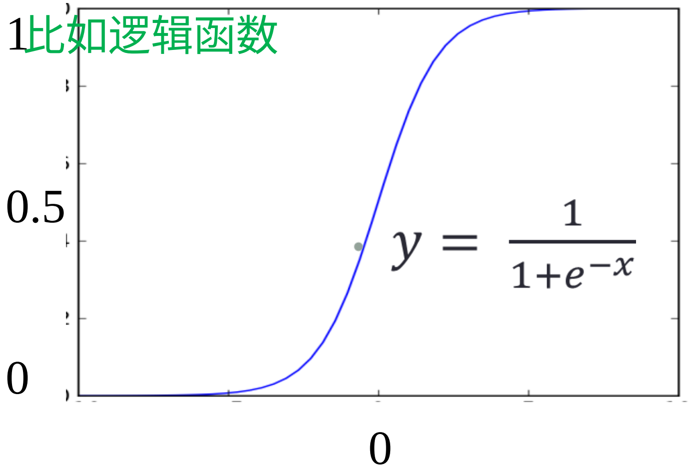


- Problem: 问题：线性回归预测y在正负无穷之间，但概率p在01之间
  - linear regression predicts *y* between –∞ and +∞
  - but probability is always between 0 and 1!

- Solution: 解决方案：我们信息网概率是通过一个模型来估计的，比如逻辑函数
  - we want our probabilities to be estimated by a model such as the logistic function
  - why? Because whatever x, it will always return a value between 0 and 1

因为这样无论x是什么
他总返回一个01
之间的值

$$y = \frac{1}{1+e^{-x}} = \frac{1}{1+e^{1/x}}$$



链接函数

Link function: Logit

- We need a *link function* f to be the x in the logistic function $y = \frac{1}{1+e^{-x}}$ and calculate y as probability p:

将链接函数f作为x，将y当做概率p

$$p = \frac{1}{1+e^{-f}} = \frac{1}{1+\frac{1}{e^f}}$$

- But $p = \frac{1}{1+\frac{1}{odds}}$

- Therefore e^f = odds; or f = log(odds)

- The link function f is called **logit p**:

  f =logit p = log(odds) = $\log(\frac{p}{1-p})$

another derivation:

- If we want $p = \frac{1}{1+e^{-f}}$, then:
- $p = \frac{e^f}{e^f+1}$
- $p(e^f+1) = e^f$
- $pe^f + p = e^f$
- $p = e^f - pe^f$
- $p = e^f(1 - p)$
- $e^f = \frac{p}{1-p}$
- $\log(e^f) = \log(\frac{p}{1-p})$
- $\boldsymbol{f = \log(\frac{p}{1-p})}$
- note: logit is always natural log (i.e. log on base e=2.71)

f = logit = log(odds of event)

范围正负无穷

- f = logit or log(odds) range from $-\infty$ to $+\infty$
  - therefore we can predict logits with a linear regression on our X_1, X_2 etc. variables

因此我们可以通过X1X2等变量的线性回归来预测logits

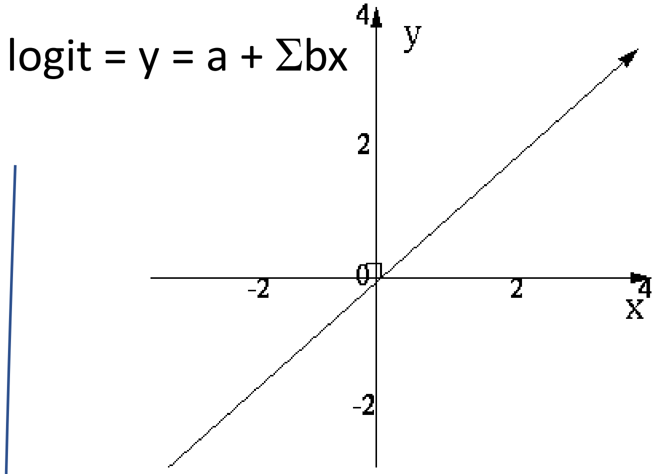
- The *logistic regression model* is thus

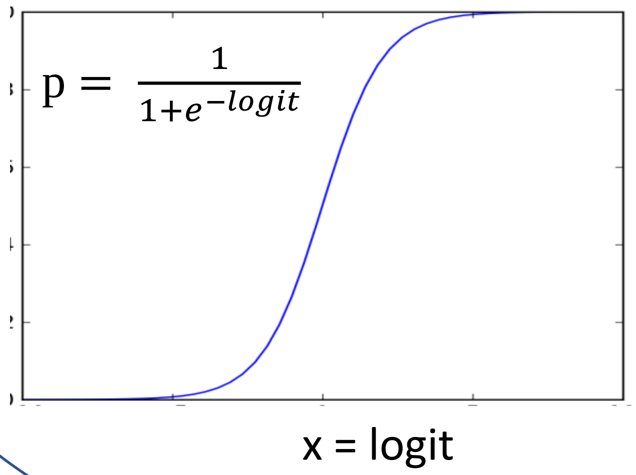
$$f = \text{logit} = \log\left(\frac{p}{1-p}\right) = \log(\text{odds}) = a + \Sigma b_i X_i$$

然后用逻辑回归中的logit来获得0和1之间的概率

- Then we use the logits in the logistic regression to obtain probabilities between 0 and 1 !

$$p = \frac{1}{1+e^{-f}} = \frac{1}{1+e^{-\log(odds)}}$$





拟合逻辑回归

Fitting logistic regression

参数a和b市通过MML方法估计的，不是通过最小二乘法

- The parameters *a* and b_i are estimated by MML (method of maximum likelihood), not by least squares
  - (we can't expand on MML in this course)

因此，统计显著性和拟合度不是基于方差，而是基于观测值和预测值之间“ 偏差 ”的度量
=个体案例预测正误的对比

- For this reason, statistical significance and goodness of fit are based not on variance, but on measures of 'deviance' between observed and predicted values
  - = comparison between right and wrong predictions of individual cases

但在线性回归中，估计的参数(系数、截距)具有决定其显著性的p值
基于与t和正态分布相关的z分布的显著性检验，解释为t-test或者F-test

- But as in linear regression, estimated parameters (coefficients, intercept) have a *P*-value that determines their significance
  - significance test based on a *z*-distribution related to *t* and normal distributions
  - interpreted like *t*-tests or *F*-tests. i.e. parameter is significant if $P<0.05$; 95% confidence intervals are provided etc. 举例：如果p<0.05则参数显著，提供95%置信区间等

逻辑回归：分类变量

Logistic regression: categorical variable

Example: let's say we want to test the effect of smoking (X, yes or no) on hypertension (Y, also yes or no)

- Y=0: no hypertension; Y=1: hypertension Y=0无高血压，Y=1高血压 X=0非吸烟者(基线组)
- X=0: non-smoker (baseline group); X=1: smoker (exposure group) X=1吸烟者(暴露组)

逻辑回归的模型是

- Logistic regression model is then:

f = logit p = log(odds of hypertension) = a + b*X

In baseline group, X=0; Therefore

- **log(odds of having hypertension when X=0) = a + b*0 = a**

=the intercept = baseline = reference level (that is, the level of hypertension for non-smokers X = 0) 截距=基线=参考水平

对a取幂，得到基线的概率

If we exponentiate a, we obtain odds at baseline

- e^a = odds of hypertension for non-smokers
- $p = \frac{1}{1+e^{-a}} = \frac{odds\ of\ non-smokers}{1+odds\ of\ non-smokers}$ = probability of hypertension for non-smokers

这些是基线值

- Those are the ***baseline values***, i.e. the odds and probabilities for groups without exposure (when all X_i=0, i.e. even if nobody smoked) a+b=log(暴露组的概率)

a + b = log (odds in the exposure group)

- Now the odds for smokers:
  - f = logit = a + bX = a + b.1 = a + b

a + b = log(odds of hypertension for smokers)

$e^{a+b} = e^a e^b$ = the odds of hypertension for smokers

$$p = \frac{1}{1+e^{-(a+b)}} = \frac{odds\ of\ smokers}{1+odds\ of\ smokers}$$ = probability of hypertension for smokers

Those are the results for the ***exposure group*** (smokers)

这些是暴露组的结果

Important: b=log(odds ratio)

So what is b then?

How likely is hypertension if you are a smoker compared a non-smoker?

- answer: it is the odds ratio (of hypertension in smokers to non-smokers)!

So: b=log(概率比)

odds ratio = odds(hypert. in smokers)/odds(hypert. in non-smokers)= $e^a e^b / e^a = e^b$

And:

log(odds ratio) = log(e^b)=b

逻辑回归中的系数b是log(暴露组中的高血压相对于基线)

- **The coefficient *b* in the logistic regression is the log(odds of hypertension in exposure group *relative to baseline*)**
  - in logistic regression, we test for significance of coefficient *b*
    - same as in linear regression! 逻辑回归中检验b的显著性和线性回归相同，对于变量的显著性，需要b不同于0，p<0.05
    - for a significant effect of variable, we need *b* different from 0 (i.e. P value < 0.05)
  - if b=0 (non-significant) 如果b=0，则不显著，暴露/基线=1，对于暴露和基线概率是相同的 =变量x对事件y的概率是没有影响的
    - odds ratio for exposure vs. baseline = $e^b = e^0 = 1$
    - = the odds are the same for exposure and baseline (1 to1),
    - = the variable X has no effect on probability of event Y

Odds ratio b

- Let's add some ***hypothetical*** numbers to the example:

  - odds of hypertension for smokers ($=e^{a+b}$) = 0.3
  - odds of hypertension for non-smokers ($=e^{a}$) = 0.1

- The odds of hypertension in smokers would be three times higher in smokers
  - ***odds ratio*** = odds smokers/odds non smokers = 0.3/0.1 = 3

两组的概率比是一个非常有用的指标，代表了一个因素对事件发生的影响

- The ***odds ratio of the two groups (exposure/baseline)*** is a very useful representation of the effect of a factor on the occurrence of event

逻辑回归中总是给b或log，不是直接给事件的比值

- Logistic regression always reports b or **log of odds,** not odds of event in exposure group relative to baseline
  - more precisely, R reports ***log(odds ratio of event in exposure vs. baseline)***
  - so in this example above, R would report b=log(3)=1.098612
    - We have to exponentiate b to obtain odds ratio = 3

R中总是给log值，需要自己取幂才能得到概率比

Break

Example in R: hypertension, smoking, obesity

- File *hypertension* presents data on people with or without hypertension as a function of two factors: smoking and obesity

- Cases coded as ‘yes’ or ‘no’ no按字母顺序排在第一位，并作为基线读取
  - ‘no’ comes first alphabetically and is read as baseline
  - alternatively: ‘no’=0, ‘yes’=1 (don’t use 1 or 2!!!) no=0 yes=1

  - In this example, data are presented as a table
    - (we’ll see a different way of presenting data with each case as a line)

>hypertension

	smoking	obesity	total	hyper	nonhyper
1	no	no	247	40	207
2	yes	no	102	15	87
3	no	yes	59	16	43
4	yes	yes	25	8	17

Example in R: hypertension, smoking, obesity

- When data are presented as table 表中有高血压和非高血压的数量
  - table has number of positives (hypertension, Y = 1) and negatives (no hypertension, Y = 0) 两个预测变量：吸烟和肥胖
  - two predictors or X variables: X_1 = smoking, X_2 = obesity
    - For both, yes = 1, no = 0
  - this has been done already for you (file *hypnonhyp*)
    - i.e. the dependent variable will be the matrix *hypnonhyp*

- Row 1: non-smoker, non-obese
- Row 2: smoker, non-obese
- Row 3: non-smoker, obese
- Row 4: smoker, obese

	hyper	nonhyper
1	40	207
2	15	87
3	16	43
4	8	17

Note: don’t worry about the table!

- You will not be asked to create one!

使用glm
线性回归是lm

Running model

```
> model.hyper <- glm(hypnonhyp ~ smoking+obesity, binomial)
> summary(model.hyper)
Call:
glm(formula = hypnonhyp ~ smoking + obesity, family = binomial)
Deviance Residuals:
      1       2       3       4
 0.1593 -0.2520 -0.2653  0.4018
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.67143    0.16731  -9.990  < 2e-16 ***
smokingyes  -0.01654    0.27617  -0.060  0.95224
obesityyes   0.76005    0.28270   2.689  0.00718 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 7.15022  on 3  degrees of freedom
Residual deviance: 0.32067  on 1  degrees of freedom
AIC: 23.935
Number of Fisher Scoring iterations: 3
```

否则更符合线性而不是逻辑回归

- Logistic regression is an example of generalised linear model
  - function *glm*
- Syntax is simple and similar to linear regression

逻辑模型写的像有两个预测因子的多元回归

- Logistic model written like a multiple regression with *two* predictors:
  - *hypnonhyp ~ smoking+ obesity*
  - (ps. interactions later)

自变量一定要添加binomial

- Argument *binomial* sets logistic regression
  - Never forget to add:

  family = binomial

  Otherwise it fits a linear rather than the logistic regression!!!

残差

Residuals

```
> model.hyper <- glm(hypnonhyp ~ smoking+obesity, binomial)
> summary(model.hyper)
Call:
glm(formula = hypnonhyp ~ smoking + obesity, family = binomial)
Deviance Residuals:
      1         2         3         4
 0.1593  -0.2520  -0.2653   0.4018
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.67143    0.16731  -9.990  < 2e-16 ***
smokingyes   -0.01654    0.27617  -0.060  0.95224
obesityyes    0.76005    0.28270   2.689  0.00718 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 7.15022  on 3  degrees of freedom
Residual deviance: 0.32067  on 1  degrees of freedom
AIC: 23.935
Number of Fisher Scoring iterations: 3
```

残差是作为偏差(而不是方差)给出的

- Residuals are given as deviance (not variance) 每组中观察到的和预测的logit值之间的差异(预测的就是上面表格的yesno)
  - difference between observed and predicted logit values in each group (no/no, no/yes, yes/no, yes/yes)
  - residuals in logit scale (neither probability nor cell count)

截距

Intercept

```
> model.hyper <- glm(hypnonhyp ~ smoking+obesity, binomial)
> summary(model.hyper)
Call:
glm(formula = hypnonhyp ~ smoking + obesity, family = binomial)
Deviance Residuals:
      1        2        3        4
 0.1593  -0.2520  -0.2653   0.4018
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.67143    0.16731  -9.990  < 2e-16 ***
smokingyes  -0.01654    0.27617  -0.060  0.95224
obesityyes   0.76005    0.28270   2.689  0.00718 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 7.15022  on 3  degrees of freedom
Residual deviance: 0.32067  on 1  degrees of freedom
AIC: 23.935
Number of Fisher Scoring iterations: 3
```

截距a=-1.67

- **Intercept a = -1.67**

a=log(基线组，高血压概率)

- a=log(odds of hypertension, baseline group)
  - =non-smokers (X1=0), non-obese (X2= 0)
  - e^a =the odds of hypertension if you're non-smoker, non-obese
  - a=0.188

a取幂等于不吸烟不肥胖，得高血压的概率

z检验：截距明显不同于0

基线中的高血压概率不是1

样本中基线时高血压概率不是0.5

- z-test: intercept is significantly different from 0
  - odds of hypertension in baseline (e^a)= not 1
  - probability of hypertension in baseline different from 0.5 in the sample

Effect of smoking

```
> model.hyper <- glm(hypnonhyp ~ smoking+obesity, binomial)
> summary(model.hyper)
Call:
glm(formula = hypnonhyp ~ smoking + obesity, family = binomial)
Deviance Residuals:
      1        2        3        4
 0.1593  -0.2520  -0.2653   0.4018
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.67143    0.16731  -9.990  < 2e-16 ***
smokingyes    -0.01654    0.27617  -0.060  0.95224
obesityyes     0.76005    0.28270   2.689  0.00718 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 7.15022  on 3  degrees of freedom
Residual deviance: 0.32067  on 1  degrees of freedom
AIC: 23.935
Number of Fisher Scoring iterations: 3
```

吸烟回归系数b

- **Regression coefficient b for smoking:**
  - smokers (X=1) are shown as *smokingyes*,
    - variable name plus group ('yes')
  - b=log(odds ratio)=-0.0165
  - =log odds of hypertension for smokers relative to non-smokers

b=log(吸烟得高血压概率/不吸烟得高血压概率)

- But P(z) = 0.95! b与0无显著差异
  - b is not significantly different from 0
  - odds ratio not different from $e^0 = 1$

概率比与1无差异

所以在这个实验中吸烟者和不吸烟者得高血压概率是一样的

- So smokers are not more likely to have hypertension than non-smokers ***in this hypothetical sample***
  - (don't start smoking because of me!)

Effect of obesity

```
> model.hyper <- glm(hypnonhyp ~ smoking+obesity, binomial)
> summary(model.hyper)
Call:
glm(formula = hypnonhyp ~ smoking + obesity, family = binomial)
Deviance Residuals:
      1        2        3        4
 0.1593  -0.2520  -0.2653   0.4018
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.67143    0.16731  -9.990  < 2e-16 ***
smokingyes       -0.01654    0.27617  -0.060  0.95224
obesityyes        0.76005    0.28270   2.689  0.00718 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 7.15022  on 3  degrees of freedom
Residual deviance: 0.32067  on 1  degrees of freedom
AIC: 23.935
Number of Fisher Scoring iterations: 3
```

在这个样本中
肥胖使高血压的几率
增加了一倍多

肥胖的回归系数b=0.76

- **Regression coefficient b for obesity: b=0.76**
  - =log odds of hypertension for obese relative to non-obese
- P(z) = 0.00718 b明显不同于0
  - b is significantly different from 0
  - b = ln(odds of hypertension in obese relative to baseline) > 0
  - odds ratio= $e^{0.76}$ =2.14 概率比2.14
    - odds ratio >1; obese at higher risk!

概率比>1，肥胖有危险

- Obesity more than doubles odds of hypertension ***in this sample***

拟合度

Goodness of fit

```
> model.hyper <- glm(hypnonhyp ~ smoking+obesity, binomial)
> summary(model.hyper)
Call:
glm(formula = hypnonhyp ~ smoking + obesity, family = binomial)
Deviance Residuals:
     1        2        3        4
 0.1593  -0.2520  -0.2653   0.4018
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.67143    0.16731  -9.990  < 2e-16 ***
smokingyes     -0.01654    0.27617  -0.060  0.95224
obesityyes      0.76005    0.28270   2.689  0.00718 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 7.15022  on 3  degrees of freedom
Residual deviance: 0.32067  on 1  degrees of freedom
AIC: 23.935
Number of Fisher Scoring iterations: 3
```

MML不用方差来衡量拟合度

- MML does not use variance to measure goodness of fit
  - it includes no 'dispersion parameter', which has to be taken as 1

在MML中，偏差取代方差

- In MML, deviance replaces variance 零偏差=当模型仅包括截距时的偏差(=预测因子吸烟和肥胖之前)
  - null deviance = deviance when model includes only intercept (=before predictors *smoking* and *obesity*)

剩余偏差是预测因素后无法解释的偏差

  - residual deviance is unexplained deviance after predictors

零偏差和剩余偏差之间的差异是预测因子对模型的贡献

  - difference between null and residual is the contribution of predictors to model

Goodness of fit

```
> model.hyper <- glm(hypnonhyp ~ smoking+obesity, binomial)
> summary(model.hyper)
Call:
glm(formula = hypnonhyp ~ smoking + obesity, family = binomial)
Deviance Residuals:
     1        2        3        4
 0.1593  -0.2520  -0.2653   0.4018
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.67143    0.16731  -9.990  < 2e-16 ***
smokingyes   -0.01654    0.27617  -0.060  0.95224
obesityyes    0.76005    0.28270   2.689  0.00718 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 7.15022  on 3  degrees of freedom
Residual deviance: 0.32067  on 1  degrees of freedom
AIC: 23.935
Number of Fisher Scoring iterations: 3
```

因为没有方差，所以拟合度不是用R方来衡量的
我们用AIC(akaike信息准则)来替代

- Because there is no variance, goodness of fit is not measured by R^2
  - we use AIC (Akaike Information Criterion) instead

在回归中添加额外的预测因子可能会增加拟合优度，即使预测因子并不显著

- Remember: adding additional predictors to regression may increase goodness of fit even when predictor is not significant

AIC测量拟合优度，同时惩罚使用额外预测因子的模型

- AIC measures goodness of fit while punishing models for use of additional predictors
  - *the better and more parsimonious the model, the lower the AIC*

模型越好、越节约，AIC就会越低

- Models with lowest AIC are selected

选择AIC最低的模型

解释和计算指南

Guide to interpretation and calculations:

- a = log(odds of event in baseline group)
- exp(a) = baseline odds of event
- Probability p of event in baseline: baseline odds/(1 + baseline odds)

基线事件概率：基线比值/(1+基线比值)

Then

如果b显著则不同于0

- b = log(odds ratio); if b is significant (different from 0):
- exp(b) = odds ratio

暴露组的比值

- exp(a+b) = exp(a)*exp(b) = odds(baseline)*odds ratio = odds in exposure group
- Probability p in exposure group = exposure odds/(1 + exposure odds)

```
model.hyper2 <- glm(hypnonhyp ~ obesity, binomial, data= )
```

Exercises

- Since *smoking* is not significant, you must optimise the model by excluding *smoking*, and run model only with variable *obesity* (manually, or with *step* function)

1. Is a significant? What does that mean?
2. Is b significant? What does that mean?

- Calculate:

3. Baseline odds of hypertension
4. Odds ratio of hypertension (obese vs. non-obese)
5. Odds of hypertension in obese
6. Probability of hypertension in non-obese
7. Probability of hypertension in obese