Lecture 11

Logistic regression: continuous variables

逻辑回归：连续变量

Summary: calculation of odds and probabilities

- Model with obesity only (*model.hyper2*)

Coefficients: 系数

```
> coef(model.hyper2)
(Intercept)    obesityyes
 -1.6762466   0.7599559
```

截距和b

Odds of hypertension in baseline =e^a, and odds ratio for obese = e^b

```
> exp(coef(model.hyper2))
(Intercept)  obesityyes
  0.1870748   2.1381818
```

Probability of hypertension in non-obese:

p = odds/(1+ odds) =

```
> 0.1870748/1.1870748
[1] 0.1576259
```

Odds of hypertension in obese group: e^{a+b}

肥胖组患高血压的几率

=p(baseline)*odds ratio

p(基线)*概率比

```
> 0.1870748*2.1381818
[1] 0.4001163
```

Probability of hypertension in the obese:

p = odds/(1+ odds) =

```
> 0.4001163/1.4001163
> [1] 0.2857736
```

if odds = p/(1-p)

then

p = odds/(1+odds)

连续变量

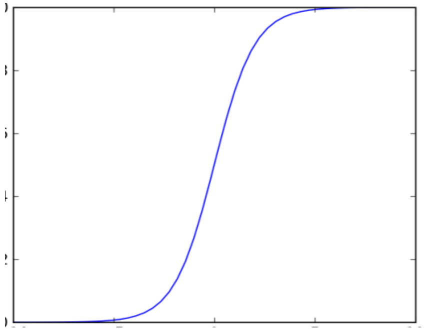
Continuous variables

我们可以古迹连续变量(如年龄)对某一事件发生概率的影响
=事件发生的累计概率

- We can estimate the effects of a continuous variable (e.g. age) on the probability of an event (e.g. menarche) having occurred
  - = the *cumulative* probability of event having occurred

如何解释连续预测因子对事件概率的影响

- How to interpret effect of continuous predictors on probability of event?



- Data: file *menar* (modified from *juul* in library *ISwR*)
  - girls aged 8 to 20 either had or haven’t had menarche (i.e. they are either ‘yes’ or ‘no’ for menarche)
    - no: menarche=0
    - yes: menarche=1

对于初潮的回答是 “ 是 ” 或者 “ 否 ”

  - logistic regression can estimate probability of menarche having occurred by age *x*

逻辑回归可估算出x岁月经初潮发生的概率

```
model.menar <- glm(menarche~age,binomial, data=menar)
```

对于初潮与年龄进行逻辑回归

Menarche and age

```
> model.menar <- glm(menarche~age,binomial)
> summary(model.menar)
Call:
glm(formula = menarche ~ age, family = binomial)
Deviance Residuals:
    Min         1Q       Median        3Q       Max
-2.32759  -0.18998   0.01253   0.12132   2.45922
Coefficients:
                  Estimate  Std. Error    z value   Pr(>|z|)
(Intercept)   -20.0132      2.0284      -9.867    <2e-16 ***
age                1.5173      0.1544       9.829    <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 719.39  on 518  degrees of freedom
Residual deviance: 200.66  on 517  degrees of freedom
AIC: 204.66
Number of Fisher Scoring iterations: 7
```

- Let's run a logistic regression of menarche against age

a=截距，在这种情况下没有实际意义，e的a次方是指0岁来月经的概率

概率为0

- a=intercept not really meaningful in this case
  - log(odds) of menarche in people with 'no age' (age=0)

b=log(初潮/无初潮 的概率比)

- b = log(odds ratio of menarche to no menarche)>1; P~0
- odds ratio= $e^b = e^{1.5176} = 4.56$
  - exposure to a 'unit of age' significantly increases odds of menarche to no menarche

增加一个 " 年龄单位 " 会明显增加初潮/无初潮的比率

Menarche and age

```
> model.menar <- glm(menarche~age,binomial)
> summary(model.menar)
Call:
glm(formula = menarche ~ age, family = binomial)
Deviance Residuals:
    Min        1Q     Median       3Q       Max
-2.32759  -0.18998   0.01253   0.12132   2.45922
Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -20.0132      2.0284   -9.867   <2e-16 ***
age            1.5173      0.1544    9.829   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 719.39  on 518  degrees of freedom
Residual deviance: 200.66  on 517  degrees of freedom
AIC: 204.66
Number of Fisher Scoring iterations: 7
```

在连续变量年龄的情况下，概率比是指当年龄增加一个单位(即每年)时事件(初潮)发生概率的变化

- In the case of continuous variable *age*, **odds ratio is change in odds of event (menarche) when age increases by 1 unit (i.e. per year)**

b值暴露于“ 一个年龄单位 ”(即一年)的影响

- *b* measures the effect of exposure to ‘one unit of age’ (i.e. one year)
  - it is the log of odds ratio of the event occurring at age N+1 to event occurring at age N

是发生在n+1岁的事件与发生在n岁的事件的概率比的对数

- =odds of menarche to no menarche increase by 4.56 per year

每年增加4.56

- 4.56 seems to be a large number, but don’t forget that the odds start at ~ 0

4.56看起来很大，但别忘了最开始是约等于0

预测概率

Predicted probabilities

对于吸烟(是或否)这样的分类预测因子，我们只需要计算两组的odd和概率：基线组和暴露组

- For categorical predictors like smoking (yes or no), we only need to calculate odds and probabilities for two groups: baseline and exposure

但对于连续变量来说，x值和y值概率都有一个范围

- But for a continuous variable, there is a range of *x* values and *y* probabilities (e.g. probabilities of menarche as a function of ages from 8 to 20)

- To predict probabilities (= $\frac{1}{1+e^{-(a+bX)}}$) of menarche for all ages from 8 to 20:
  - function *predict* 用函数predit
  - add argument *type="response"* 添加这个参数，直接给概率
    - otherwise *predict* returns logit values 否则用predit返回log值

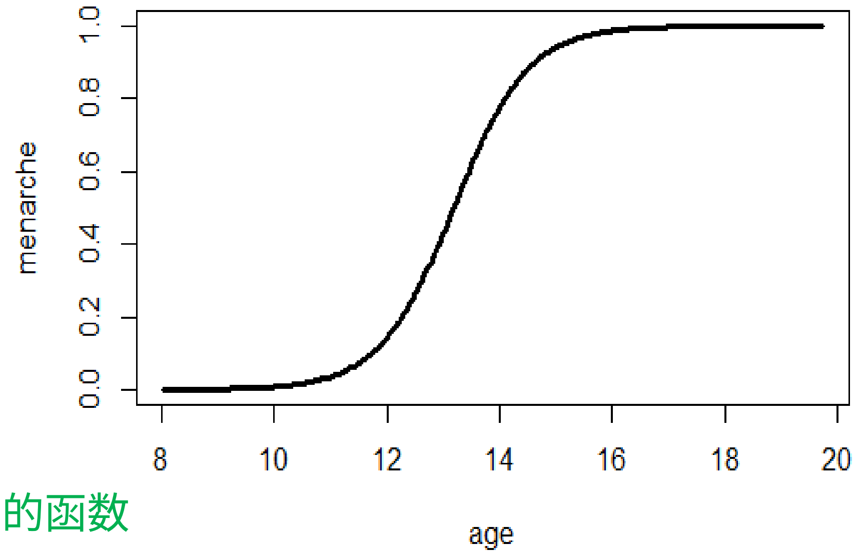
- Saving probabilities in vector *prob*: 在向量prob中保存概率

  > probs <- predict(model.menar, type="response")

按照年龄绘制初潮概率图

- Plotting probability of menarche by age

  > plot(probs~age, data=menar, pch=16, ylab="menarche")



- To predict probabilities at a given point, use 要预测给定点的概率，用下面的函数

  > predict(your model, data.frame(X = value), type= "response")

事件发生时的中位年龄

Median age at event occurrence

初潮年龄中位数=初潮发生概率为50%的年龄，或者p=0.5

- Median age at menarche = age where probability of menarche having occurred is 50%, or p=0.5

But when p=0.5:

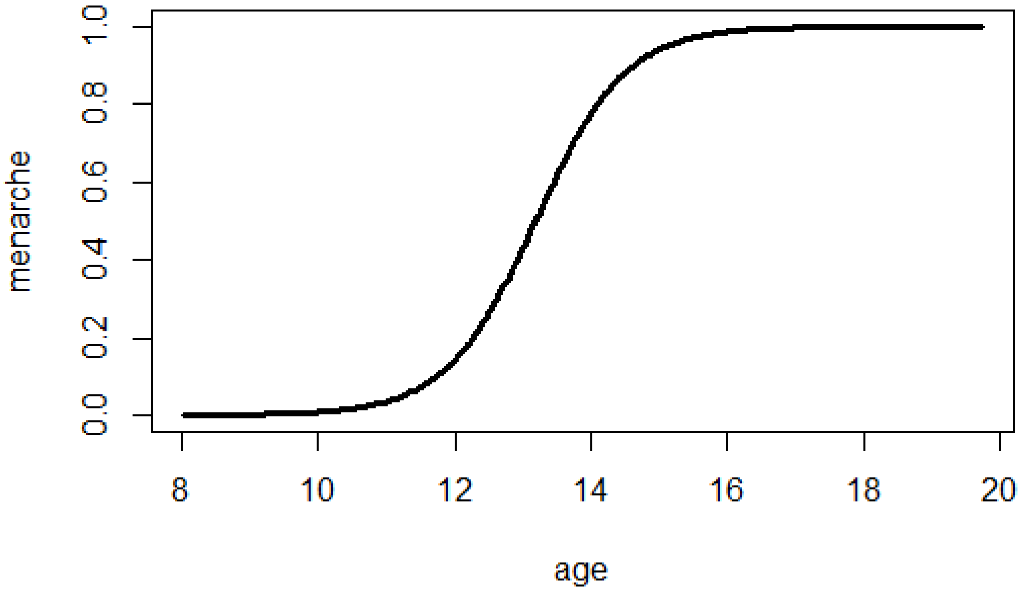
当p=0.5时，odds=1
log(odds)=0

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

If odds = 1

then

- log(odds) = 0



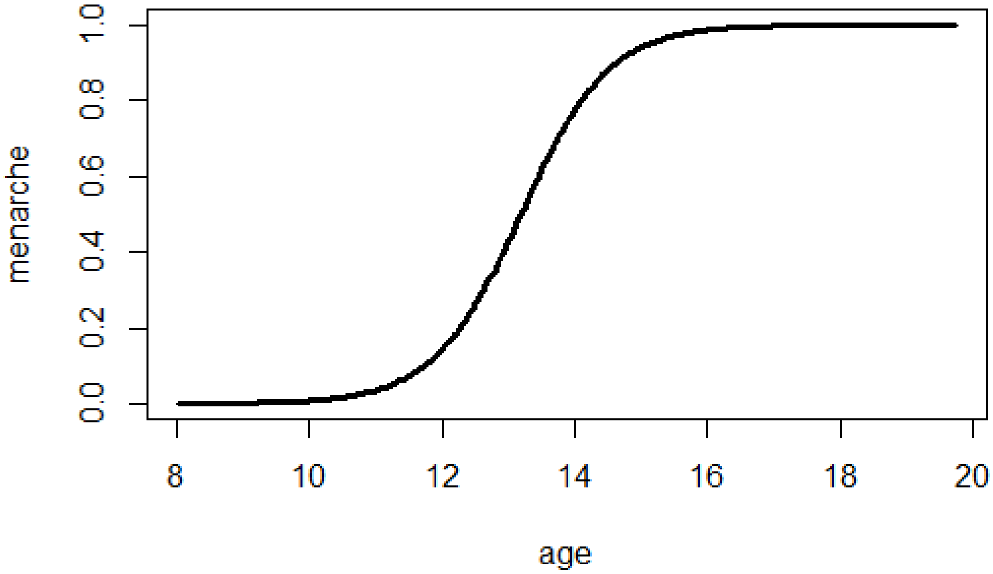
Median age at event occurrence

设log(odds)=logit=10，可以根据公式计算出初潮年龄x

- Setting log(odds) = logit = 0, we can calculate age X at menarche from the equation:
  - logit p = a + bX

  - 0 = a + b(age)
  - 0 = -20.013 + 1.5173(age)
  - **median age = 13.19 years**

中位年龄=13.19



Exercise

- Still using file *menar*, run a logistic model using *igf1* (insulin-growth-like factor 1) as predictor of menarche

1. Interpret a and b

2. What is the probability of menarche for someone with igf1=500?

Tips:

- Estimate logit p (=*f*=a + bX) when *igf1* = 500
- Then use logistic function (p = 1/(1 +exp(-f)) to convert logit p into probability
- Confirm with this code:

```
> predict(model.menar, data.frame(igf1 = 500), type= “response”)
```

Break

大于2级的分类变量

Categorical variables with > 2 levels

有些分类预测因子有两个级别(1和0)，但有些有更多级别，如月份地点等

- Some categorical predictors have two levels (smoker=1, non-smoker=0), but others have more levels (month, location etc.)

我们仍然可以利用这些变量进行逻辑回归，但解释有略微变化

- We can still run a logistic regression with those variables, but interpretation slightly changes again

例如：婴儿数据集中的月份

- Example: month in *infant* dataset

当预测因子大于2级时：第一级为基线，所有其他级别（其他月份）与第一级分别进行单独比较，而不是相互比较

- When predictor has >2 levels
    - first level is taken as baseline month=1,然后输入as.factor(month)
        - =January (coded as month=1 and then entered as *as.factor(month)*)
    - all other levels (=other months) are compared to the first *on an individual basis*, but not to each other
        - =February vs. January, March vs. January etc.
        - but not March vs. February etc.

Categorical variables with > 2 levels

结果：健康变量（0营养不良，1健康）

- Outcome: variable *healthy* (0=undernourished, 1=healthy)

预测因子：月份（出生月份）

- Predictor: *month* (birth month)
  - 9 levels (January=1 to September=9)
  - check file: month is a factor (although coded as numbers)
    - always make sure your categorical predictor is set as factor or character, NEVER numeric
    - otherwise you get a ‘continuous’ predictor (like age in the menarche analysis!)
  - “1” is baseline level “January”, “2” is level “February”

总共九级

始终确保分类预测因子设置为因子或字符
而不是数字
否则就会得到一个“ 连续 ”的预测因子
（像月经初潮分析中的年龄）

```
> model.infant <- glm(healthy ~ month,binomial, data=infant)
```

离散程度大于二级的分类变量

Categorical variables with > 2 discrete levels

```
 > model.infant <- glm(healthy ~ as.factor(month),binomial, data=infant)
> summary(model.infant)
Coefficients:
              Estimate   Std. Error z value  Pr(>|z|)
(Intercept)  -2.18407    0.25582  -8.537   <2e-16 ***
month2       -0.06723    0.36678  -0.183   0.8546
month3       -0.21383    0.37864  -0.565   0.5723
month4       -0.81167    0.44350  -1.830   0.0672 .
month5       -0.81167    0.44350  -1.830   0.0672 .
month6       -1.28167    0.52124  -2.459   0.0139 *
month7       -1.25635    0.52140  -2.410   0.0160 *
month8       -0.97293    0.48909  -1.989   0.0467 *
month9       -1.14814    0.52210  -2.199   0.0279 *

    Null deviance: 642.54  on 1457  degrees of freedom
Residual deviance: 623.81  on 1449  degrees of freedom

AIC: 641.81
```

只有6-9月与1月(基线)有明显差异，系数b明显<0

- Only months 6 to 9 (June to September) significantly differ from January (baseline)
  - coefficients b (log of odds ratio) are significantly <0

- Let’s compare January and June:
- Exponentiating coefficients: 指数化系数

```
>exp(coef(model.infant))
```

- Odds in January = exp(a)

```
0.1125828
```

- Odds ratio June to January = odds ratio month 6 to month 1

```
0.2775735
```

- Odds of being healthy in June: 六月健康的概率

=0.1125*0.277535

=0.03123487

Categorical variables with > 2 discrete levels

- Remember:

当预测因子有两个以上水平时，系数和p值反映了每个暴露水平与基线之间的比较情况

- When there are more than two levels in predictor, coefficients and P values reflect comparisons between each exposure level and baseline
    - Each month compared to January

但暴露水平之间没有可比性，如四月和五月之间

- **But there is no comparison between exposure levels**
    - We know nothing directly about the difference between April and May

如果想知道，可以将基线改成4

- If we wanted to know about April vs. May:
    - Change baseline to April = 4 (with function *relevel*), then we would obtain a coefficient for May vs. April

```
levels(infant$month)
infant$month <- relevel(infant$month, ref = "6")
```
将六月改成baseline

Exercise

- Re-run *model.infant*, but this time use month=9 (September) as baseline

(see code file)

```
levels(infant$month) <- c("1", "2", "3", "4", "5", "6", "7", "8",
 "9")
```
再改回来

```
infant$month <- relevel(infant$month, ref = "9")
```
将九月改成baseline

Break

Interactions

当各因素的影响不是独立的（他们不是相加而是相乘时）
就会产生相互作用

- Interactions occur when the effects of factors are not independent (they are not additive but multiplicative)

- Positive interaction: 积极相互作用
  - Drug A causes small increase in odds of heart attack
  - Drug B causes small increase in odds of heart attack
  - But people taking drugs A and B have a large increase in odds of heart attack

  =positive interaction between A and B: their effects are stronger when combined

  =A和B之间的正向相互作用：两者结合之后效果增强

- Negative interaction: 消极相互作用
  - Drug A causes increase in odds of heart attack
  - Drug B causes increase in odds of heart attack
  - But people taking drugs A and B show no increase in odds of heart attack

  =negative interaction between A and B: their effects are cancelled or reduced when combined

  =两者之间的负相互作用：两者结合后效果被抵消或减弱

Interactions

- Interaction occurs if 如果因素1和因素2共同出现在一个人身上时，他们的共同效应应不同于1和2的相加效应 这时就有interaction
  - factors 1 and 2 are present in the same individual
  - their joint effect is different from the additive effects of 1 and 2 put together.

Example:

- Exposure to factor A doubles odds of an outcome
- Exposure to factor B also doubles odds of the outcome

What to expect from exposure to both A and B?

=(exposure to A) x (exposure to B) = 2 x 2 = 4 times the odds

=***additive*** effect of A and B A和B的叠加效应

但如果同时接触AB会导致几率不同于4（叠加效应），则AB是相互作用的

- But if exposure to both A and B results in odds different from 4 (the additive effect), A and B are *interacting*
  - joint effect > 4: ***positive*** interaction
  - joint effect < 4: ***negative*** interaction

联合效应>4：正相互作用
联合效应<4：负相互作用

Interactions

- File *evans*: Evans county study of factors leading to coronary heart disease

- Let’s focus son the effects of three factors:
  - *age*
  - *cat* (catecholamine levels) and
  - *chl* (cholesterol levels)

on the probability of coronary heart disease (*chd)*

三个变量：
年龄
儿茶酚胺水平
胆固醇水平
对于冠心病发病的概率

在R中，包含变量X1和X2之间所有可能得相互作用，将他们相乘X1*X2

- In R, to include all possible interactions between variables X1 and X2,
  - multiply them: Y ~ X1*X2
- X1*X2 is expanded into:
  - X1 + X2 + X1:X2
  - interactions are represented by “:”

X1*X2被扩展成：X1+X2+X1X2，相互作用用“ ： ”表示

Our model with interactions is then:

模型为

```
> model.chd <- glm(chd ~ age*cat*chl, binomial, data=evans)
```

基线

Interactions: baseline

```
> model.chd <- glm(chd~age*cat*chl,binomial,data=evans)
> summary(model.chd)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3268  -1.1954   0.8112   1.1154   1.6543
 Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.6564566  3.1060236  -1.821  0.06859 .
age          0.0929091  0.0589832   1.575  0.11522
cat         28.3920812 10.9032473   2.604  0.00921 **
chl          0.0223684  0.0140188   1.596  0.11058
age:cat     -0.5281193  0.1861421  -2.837  0.00455 **
age:chl     -0.0003483  0.0002650  -1.314  0.18873
cat:chl     -0.1302252  0.0546123  -2.385  0.01710 *
age:cat:chl  0.0024763  0.0009319   2.657  0.00788 **

Null deviance: 840.31  on 608  degrees of freedom
Residual deviance: 809.76  on 601  degrees of freedom
AIC: 825.76
```

- First let’s understand how to read interactions:

cat会增加患冠心病的几率b=28.39，年龄的影响不大

- *cat* increases odds of coronary disease
  - b = 28.39
- *age* does not have a significant effect
- however, *age* and *cat* show a significant and negative interaction
  - b= -0.52

但两者的相互作用呈现显著负相关b=-0.52

Example: interaction *age:cat*

```
> model.chd <- glm(chd~age*cat*chl,binomial,data=evans)
> summary(model.chd)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3268  -1.1954   0.8112   1.1154   1.6543
 Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.6564566  3.1060236  -1.821  0.06859 .
age          0.0929091  0.0589832   1.575  0.11522
cat         28.3920812 10.9032473   2.604  0.00921 **
chl          0.0223684  0.0140188   1.596  0.11058
age:cat     -0.5281193  0.1861421  -2.837  0.00455 **
age:chl     -0.0003483  0.0002650  -1.314  0.18873
cat:chl     -0.1302252  0.0546123  -2.385  0.01710 *
age:cat:chl  0.0024763  0.0009319   2.657  0.00788 **

Null deviance: 840.31  on 608  degrees of freedom
Residual deviance: 809.76  on 601  degrees of freedom
AIC: 825.76
```

换句话说，交互项的基线不是截距a，而是a+b(age)+b(cat)

- **Important**:

cat的效果是相对于基线进行测量的，计算时我们使用截距a和cat的系数b

- Effect of *cat* is measured relative to baseline
  - to calculate it, we use intercept *a* and *cat* coefficient *b*

相比之下，age:cat这个交互作用有了一个不同的基线

- By contrast, the interaction effect *age:cat* now has a different baseline: a person where *age* and *cat* had their additive effects calculated

一个人的age和cat的相加效应被计算出来

- In other words, baseline for interaction term is not intercept a, but
  - a + b(age) + b(cat)

Example: interaction *age:cat*

```
> model.chd <- glm(chd~age*cat*chl,binomial,data=evans)
> summary(model.chd)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3268  -1.1954   0.8112   1.1154   1.6543
 Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.6564566  3.1060236  -1.821  0.06859 .
age          0.0929091  0.0589832   1.575  0.11522
cat         28.3920812 10.9032473   2.604  0.00921 **
chl          0.0223684  0.0140188   1.596  0.11058
age:cat     -0.5281193  0.1861421  -2.837  0.00455 **
age:chl     -0.0003483  0.0002650  -1.314  0.18873
cat:chl     -0.1302252  0.0546123  -2.385  0.01710 *
age:cat:chl  0.0024763  0.0009319   2.657  0.00788 **

Null deviance: 840.31  on 608  degrees of freedom
Residual deviance: 809.76  on 601  degrees of freedom
AIC: 825.76
```

交互项是在4个相加效应之外的3个效应=7个总效应

- **Important**:

- Back to our hypothetical example: if
  - effect of A = 2
  - effect of B = 2

  addictive effects: 2 x 2 = 4

- then if interaction A:B = 3
  - total effect = 2 x 2 x 3 = 12

A:B是3，小于相加效果4，这并不意味着是消极相互影响
不是减少

- A:B is 3; less than the additive effects of 4
  - this doesn't mean a negative interaction! Not a reduction!
  - interaction term is an effect of 3 on top of the additive effects of 4 = total effect of 7

Example: interaction *age:cat*

```
> model.chd <- glm(chd~age*cat*chl,binomial,data=evans)
> summary(model.chd)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3268  -1.1954   0.8112   1.1154   1.6543
 Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.6564566  3.1060236  -1.821  0.06859 .
age          0.0929091  0.0589832   1.575  0.11522
cat         28.3920812 10.9032473   2.604  0.00921 **
chl          0.0223684  0.0140188   1.596  0.11058
age:cat     -0.5281193  0.1861421  -2.837  0.00455 **
age:chl     -0.0003483  0.0002650  -1.314  0.18873
cat:chl     -0.1302252  0.0546123  -2.385  0.01710 *
age:cat:chl  0.0024763  0.0009319   2.657  0.00788 **

Null deviance: 840.31  on 608  degrees of freedom
Residual deviance: 809.76  on 601  degrees of freedom
AIC: 825.76
```

- **Important**:

对数形式的相加效应

- In log form (where multiplications become additions):
  - effect of A = 0.5
  - effect of B = 0.5
  - additive effects: A + B = 1

- A:B= 0.1: positive interaction
  - Total effect = 0.5 + 0.5 + 0.1 = 1.1

Example: interaction *age:cat*

```
> model.chd <- glm(chd~age*cat*chl,binomial,data=evans)
> summary(model.chd)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3268  -1.1954   0.8112   1.1154   1.6543
 Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.6564566  3.1060236  -1.821  0.06859 .
age          0.0929091  0.0589832   1.575  0.11522
cat         28.3920812 10.9032473   2.604  0.00921 **
chl          0.0223684  0.0140188   1.596  0.11058
age:cat     -0.5281193  0.1861421  -2.837  0.00455 **
age:chl     -0.0003483  0.0002650  -1.314  0.18873
cat:chl     -0.1302252  0.0546123  -2.385  0.01710 *
age:cat:chl  0.0024763  0.0009319   2.657  0.00788 **

Null deviance: 840.31  on 608  degrees of freedom
Residual deviance: 809.76  on 601  degrees of freedom
AIC: 825.76
```

- Now back to *chd* model

- *age:cat* interaction: b=-0.528 = negative interaction; 负相互作用

但总效应仍然是正的

- But total effect on odds of chd is still positive:

age + *cat* + *age:cat* =

0.09 + 28.39 – 0.528 = 27.952

- In summary, the effects of *cat* and *age* are reduced when they act together
  - but they still increase odds of *chd* because of the strong additive effect of *cat*

总之，当两者一起时影响会减少，但他们仍会增加得chd的概率，因为cat的强烈叠加效应

Break

模型优化和层次原理　当回归返回项不显著时，我们必须优化模型，以获得最小的适当模型

Model optimisation and the Hierarchy Principle

- When regressions return non-significant terms, we must optimise models to obtain a *minimal adequate model*

优化必须遵循分层程序：先测试高阶交互作用，最后测试单个因素 如果交互作用显著，则必须保留所有低层次的交互作用和单项，即使他们并不显著

- Optimisation must follow a hierarchical procedure: higher-order interactions are tested first, individual factors last
  - if an interaction is significant, all lower level interactions and single terms must be kept *even if they are not significant*

例如：如果交互作用XA:X2:X3显著，则最终模型要包括他们低层次的交互作用X1:X2 X1:X3 X2:X3和单项X1 X2 X3

- EXAMPLE: if an interaction *X1:X2:X3*, is significant, final model must also include
  - all their lower interactions *X1:X2*, *X1:X3*, *X2:X3*
  - its single terms *X1*, *X2*, *X3*

因为我们需要较低的水平做为基线来估算总效应=同样的原因我们需要基线中的概率来估算暴露组中的概率

- Why? Because we need lower levels as baselines to estimate total effects
  = same reason we need odds in baseline to estimate odds in exposure group

等级原则

The Hierarchy Principle

```
> model.chd <- glm(chd~age*cat*chl,binomial,data=evans)
> summary(model.chd)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3268  -1.1954   0.8112   1.1154   1.6543
 Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.6564566  3.1060236  -1.821  0.06859 .
age          0.0929091  0.0589832   1.575  0.11522
cat         28.3920812 10.9032473   2.604  0.00921 **
chl          0.0223684  0.0140188   1.596  0.11058
age:cat     -0.5281193  0.1861421  -2.837  0.00455 **
age:chl     -0.0003483  0.0002650  -1.314  0.18873
cat:chl     -0.1302252  0.0546123  -2.385  0.01710 *
age:cat:chl  0.0024763  0.0009319   2.657  0.00788 **

Null deviance: 840.31  on 608  degrees of freedom
Residual deviance: 809.76  on 601  degrees of freedom
AIC: 825.76
```

age:cat:chl是显著的

- Back to our model: triple interaction *age:cat:chl* is significant
  - (unfortunately!)

因此如果我们对这个模型进行优化，就无法舍弃任何不重要的项

- Therefore, if we optimised this model, we would not be able to discard any non-significant terms

The Hierarchy Principle

在计算总效应时，需要：

```
> model.chd <- glm(chd~age*cat*chl,binomial,data=evans)
> summary(model.chd)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3268  -1.1954   0.8112   1.1154   1.6543
 Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.6564566  3.1060236  -1.821  0.06859 .
age          0.0929091  0.0589832   1.575  0.11522
cat         28.3920812 10.9032473   2.604  0.00921 **
chl          0.0223684  0.0140188   1.596  0.11058
age:cat     -0.5281193  0.1861421  -2.837  0.00455 **
age:chl     -0.0003483  0.0002650  -1.314  0.18873
cat:chl     -0.1302252  0.0546123  -2.385  0.01710 *
age:cat:chl  0.0024763  0.0009319   2.657  0.00788 **

Null deviance: 840.31  on 608  degrees of freedom
Residual deviance: 809.76  on 601  degrees of freedom
AIC: 825.76
```

- When calculating the total effect of *age* and *cat and chl* on chd, we need:
  - intercept a
  - b_1 for *age*
  - b_2 for cat
  - b_3 for chl
  - b_{12} for age:cat
  - b_{13} for age:chl
  - b_{23} for cat:chl
  - $b_{123}$for age:cat:chl

logit *chd* = a + b_1(*age*) +b_2(*cat*)+b_{12}(*age:cat*)
= a+ b_1*X1+ b_2*X2+ b_{12}*X1*X2…etc

Example: for subject id=51, *age*=56, *cat*=1, and chl=201, logit chd
=-5.66+0.09*(56)+28.4*(1)+0.02*(201)…etc

模型优化：step()函数

Model optimisation: function *step( )*

=摒弃不必要的变量 该函数遵循分层原则

- We optimise models (=discarding unnecessary variables) using the function *step*, which follows the hierarchical principle 该函数运行ANOVA，比较包含和不包含给定项的模型，如果不包含给定项
  - *step* runs ANOVAs comparing models with and without a given term; if absence of term does not significantly change the model, it should be eliminated 不会显著改变模型，则应将其删除

优化是基于log-likehood和aic的

- Optimisation is based on the log-likelihood and AIC (Akaike information criterion, a function both of significance and number of variables in a model)
  - AIC comparisons only work for models that are hierarchically organised, i.e. when variables in model 1 are a subset of variables in model 2

AIC比较只适用于分层组织的模型，即模型1中的变量是模型2中变量的子集时

- In practical terms: 具体而言 如果变量使log-likelihood升高使AIC降低，我们就该剔除该变量
  - we eliminate a variable if this increases log-likelihood and reduces AIC
  - we test variables according to the hierarchical principle (higher-interactions first, single terms last)

我们根据分层原则对变量进行检验（先检验高级交互作用，后检验单项作用）

Example: menarche

```
> summary(model.menar2)
Call:
glm(formula = menarche ~ age * igf1, family = binomial, data =
menar)

Deviance Residuals:
    Min       1Q    Median       3Q      Max
-2.41072  -0.03565   0.01761   0.09315   2.60345

Coefficients:
                      Estimate    Std. Error   z value      Pr(>|z|)
(Intercept)           -3.162e+01 1.021e+01 -3.096       0.00196 **
age                   2.100e+00  7.633e-01  2.752       0.00593 **
igf1                  1.794e-02  1.996e-02  0.899       0.36886
age:igf1              7.769e-04  1.522e-03  -0.511      0.60962
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
   Null deviance: 564.83  on 410  degrees of freedom
Residual deviance: 111.44  on 407  degrees of freedom
  (108 observations deleted due to missingness)
AIC: 119.44
Number of Fisher Scoring iterations: 8
```

- Now let us run a model of menarche on *age* and *igf1* with interactions,

这两者都不显著，都应该从优化模型中删除

- Neither *igf1* nor *age:igf1* is significant

- They should be both eliminated from the optimised model

Example

```
> summary(step(model.menar2)) **

Call:
glm(formula = menarche ~ age + igf1, family = binomial, data =
menar)

Deviance Residuals:
    Min       1Q    Median       3Q       Max
-2.43884  -0.04581   0.01931   0.09146   2.58392

Coefficients:
              Estimate   Std. Error  z value      Pr(>|z|)
(Intercept) -26.887594 3.650184 -7.366        1.76e-13 ***
age            1.739611   0.238325   7.299        2.89e-13 ***
igf1           0.007814   0.001880   4.157        3.23e-05 ***
```

重要：删除一个项时，所有剩余项的p值都有可能发生改变
因此只有在优化之后才知道哪个是重要的
在上一页中，igfl并不显著，但剔除掉age:igfl之后他就变得显著了
只使用最终优化模型中的系数
在本例中，只有age和igfl显著

使用函数step，得到的输出比这更多

- We optimise model with function *step*

** we obtain more output than this; see code

IMPORTANT

- When you remove a term, P values of all remaining terms may change!
- Therefore you can only tell which term is significant after optimisation
  - in the full model (previous slide), *igf1* was not significant, but once *age:igf1* was removed, it became significant in the final model (this slide)
- Only use coefficients from the final, optimised model!
- In this example, only age and igf1 are significant

```
model.chd2 <- glm(chd~age*cat,binomial, data=evans)
summary(model.chd2)
summary(step(model.chd2))
```

Exercise

- Run a logistic regression of *chd* on *age*, *cat* and their interaction.
- What is the optimal model and its AIC?