Lecture 11

Logistic regression: continuous variables

Summary: calculation of odds and probabilities

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

Coefficients:

> coef(model.hyper2)

(Intercept) obesityyes

-1.6762466 0.7599559

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: = $\frac{1}{1+e^{-(a)}}$

> 1/(1+exp(-(-1.676)))

[1] 0.1576259

Or more simply: p = odds/(1+ odds) = 0.18/1.18 *

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

> exp(-1.676+0.76)

[1] 0.4001163

Probability of hypertension in the obese: $\frac{1}{1+e^{-(a+b)}}$

> 1/(1+exp(-(-1.676+0.76)))

[1] 0.2857736

Or more simply: p = odds/(1+ odds) = 0.40/(1.40)

if

odds = p/(1-p)

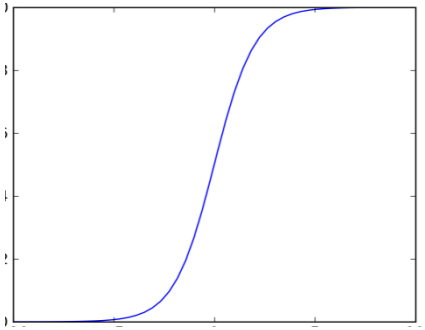
then

p = odds/(1+odds)

Continuous variables

- We can estimate the effects of a continuous variable (e.g. age) on the probability of an event (e.g. menarche) having already occurred
  - = the *cumulative* 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
  - logistic regression can estimate probability (between 0 and 1) of menarche having occurred by age



Menarche and age

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

- How do we interpret output in the case of continuous *age*?
- a=intercept is not really meaningful
  - log(odds) of menarche in people with 'no age' (age=0)
- b = ln(odds ratio of menarche to no menarche)>1; P~0
  - age significantly increases odds of menarche to no menarche
- odds ratio= $e^b = e^{1.5176} = 4.56$
  - interpretation is slightly different from discrete case
  - now exposure is age (like 'obesity')
  - so what is age=1 vs. age=0 (or more generally, age=X vs. age=X-1)?

Menarche and age

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

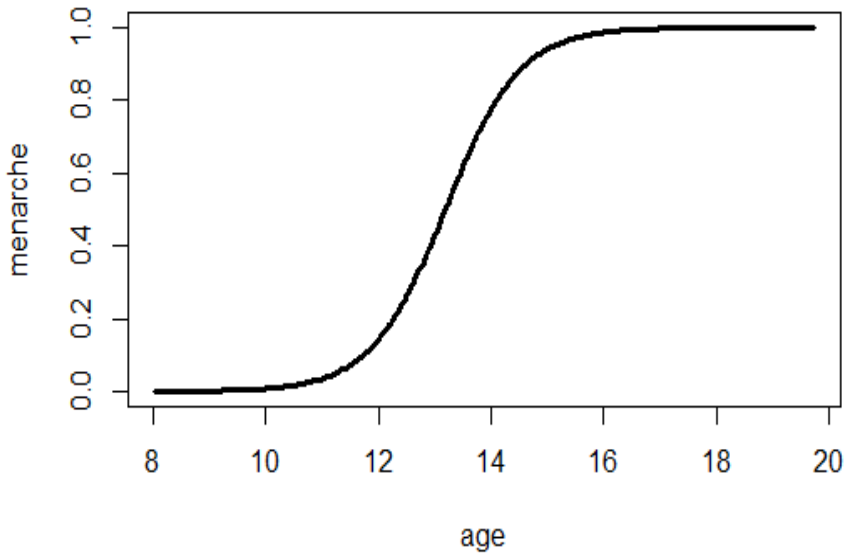
- In the case of continuous variable *age*, **odds ratio is change in odds of event (menarche) when age increases by 1 (i.e. per year)**
  - *b* assesses effect of exposure to 'one unit of age' (i.e. one year)
  - odds of menarche to no menarche (p/1-p) increase 4.56 per year (unit increase in age)
  - 4.56 seems to be a large number, but don't forget that the odds start at ~0

Predicted probabilities

- For categorical predictors, we only need to calculate odds and probabilities for two groups: baseline and exposure
- But for a continuous variable, there is a range of *x* values and *y* probabilities (e.g. probabilities of menarche as a function of age from 8 to 20)
- To predict probabilities (= $\frac{1}{1+e^{-(a+bX)}}$) of

menarche for all ages from 8 to 20:

  - function *predict*
  - add argument *type= "response"*
    - otherwise *predict* returns logit values



- Saving probabilities in vector *prob*:
  > probs <- predict(model.menar, type="response")
- Plotting probability of menarche by age
  > plot(probs~age, data=menar, pch=16, ylab="menarche", lwd=3)
- To predict probabilities at a given point, use
  > predict(your model, data.frame(X = value), type= "response")

Median age at event occurrence

- Median age at menarche = age where probability of menarche having occurred is 50%)
  - If logit p = $\ln(\frac{p}{1-p}) = 0$

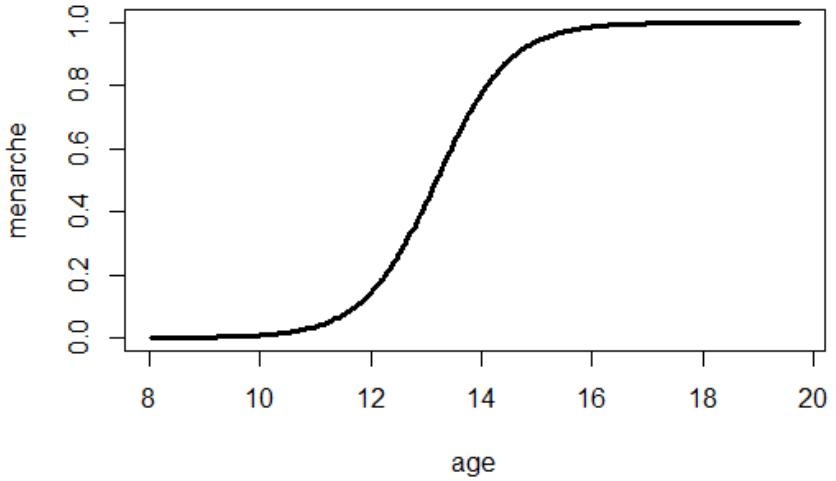
$\frac{p}{1-p} = e^0 = 1$

=> p=0.5

- i.e. when logit p =0, probability of event is 50%
  - p=0.5 is the predicted median,
    - the age at which 50% of menarches are predicted to have occurred

- Setting logitp=0, we can calculate logit from the equation:
  - logit p = a + bX

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



Exercise

- Run a logistic model using *igf1* (insulin-growth-like factor 1) as predictor of menarche

-3.239609

1. Interpret a and b

0.008647

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

Tips:

x = 500 P (500) = 1 / 1e^-(a+b*500)

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

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

Categorical variables with > 2 levels

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

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

- When predictor has >2 levels (example: month in *infant* dataset)
  - the first is taken as baseline
    - =January (coded as month=1 and then entered as *as.factor(month)*)
  - all the other levels are compared to the first on an individual basis, but not between themselves
    - =February vs. January, March vs. January etc.
    - But not March vs. February etc.

Categorical variables with > 2 levels

- Outcome: *healthy* (0=undernourished, 1=not undernourished)

- Predictor: *month* (birth month)
  - 9 levels (January=1 to September=9)
  - but we want to run month as a factor; we enter it as

    as.factor(month)
    - "1" becomes baseline level "January", "2" is level "February"

- *model.infant <- glm(healthy ~ as.factor(month),binomial, data=infant)*

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)      a  -2.18407    0.25582  -8.537  <2e-16 ***
as.factor(month)2 -0.06723  b  0.36678  -0.183  0.8546
as.factor(month)3 -0.21383    0.37864  -0.565  0.5723
as.factor(month)4 -0.81167    0.44350  -1.830  0.0672 .
as.factor(month)5 -0.81167    0.44350  -1.830  0.0672 .
as.factor(month)6 -1.28167    0.52124  -2.459  0.0139 *
as.factor(month)7 -1.25635    0.52140  -2.410  0.0160 *
as.factor(month)8 -0.97293    0.48909  -1.989  0.0467 *
as.factor(month)9 -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
```

- Baseline=January
- Only months 6 to 9 (June to September) significantly differ from January)
  - coefficients b (log of odds ratio) are significantly <0
- Let's compare January and June:

```
>exp(coef(model.infant))
(Intercept)
0.1125828

as.factor(month) 6
0.2775735
```

- Odds of malnutrition in January: 0.11258
- Odds of malnutrition in June:

=0.1125*0.277535

=0.03123487

Categorical variables with > 2 discrete levels

- Important:

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

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

- If we wanted to know about April vs. May:
  - Create a new file where April is baseline, then we would obtain a coefficient for May vs. April

Interactions

- Interactions mean that the effects of factors are not independent (they are not just additive but also multiplicative)

- Interaction occurs if
  - factors 1 and 2 are present in the same individual
  - but their joint effect is different from the sum of separate effects of 1 and 2.

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

Additive effect is expected to be 4 times more influencing

f.i. 2 or 8 times the odds, showing that A and B are interacting

- But if exposure to both A and B results in odds different from 4 (the additive effect), A and B are interacting

Interactions

- Positive interaction:
  - Drug A causes small increase in odds of heart attack
  - Drug B causes small increase in odds of hear attack
  - But people taking drugs A and B show large increase in odds of heart attack
  - =positive interaction between A and B: their effects are stronger when combined

- Negative interaction:
  - Drug A causes increase in odds of heart attack
  - Drug B causes increase in odds of hear 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

Interactions

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

- Let's examine the effects of *age*, *cat* (cathecolamine levels) and *chl* (cholesterol levels) on the probability of coronary heart disease (*chd*)

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

- In R, to include all possible interactions between variables X1 and X2,
  - Multiply them: Y ~ X1*X2
- This generates
  - X + Y + X1:X2
  - Interactions are represented by ":"

Interactions

```
> 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
```

???? How to define significance?

- Significant factors:

  P < 0.05

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

Interactions

```
> 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 note:**
- Effect of *cat* is measured relative to baseline
  - To calculate it, we use intercept *a* and act coefficient *b*
- But the interaction effect *age:cat* is relative to a hypothetical person where *age* and *cat* only had additive effects
- Baseline for interaction term is not 0
  - Baseline : a + b(age) + b(cat)

Interactions

```
> 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
```

- What does it mean?
- b for *age:cat* interaction is b=-0.52.
- This does not necessarily mean that *age:cat* people have odds of *chd* lower than baseline (i.e. a reduction of odds relative to baseline)
- it means the effects of *cat* and *age* partially reduce each other
  - but *cat* and *age* increase odds of *chd* (*cat* has positive b)

Model optimisation

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

- The way to do it is to use *anova* function to compare models with vs. without the tested term
  - if there is no significant difference, model does not need that variable

The Hierarchy Principle

- But optimisation must follow a hierarchical procedure

- The hierarchy principle means that higher-order interactions are tested first
  - If they are significant, all lower level interactions and single terms must be kept *even if they are not significant*

- In other words, if interaction *X1:X2:X3* is significant, final model must also include
  - Single terms *X1*, *X2*, *X3*
  - Interactions *X1:X2*, *X1:X3*, *X2:X3*

- The reason is that higher order interactions have coefficients that measure deviations from additive effects of lower order terms
- Therefore I need them to estimate the total effect (additive plus interactive)
  - For the same reason we need odds in baseline to estimate odds in exposure group

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
```

- In our example, triple interaction *age:cat:chl* is significant
- Therefore, optimised model must include all terms, including the non-significant ones (*age*, *chl*, *age:chl*)
- If we optimised this model, we would not discard any terms

The Hierarchy Principle

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

```
> 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
```

- Example: when calculating the effect of *age* and *cat* (on odds or probabilities), we need to use

- Intercept a
- b_1 for *age*
- b_2 for cat
- b_{12} for age:cat

logit of *chd* when *cat* increases by one unit and age increases by one unit :

$= a + b_1(age) + b_2(cat) + b_{12}(age:cat)$

$= a + b_1*X1 + b_2*X2 + b_{12}*X1*X2$

Model optimisation

- We optimise models (discarding unnecessary variables) using the function *step*, which obeys the hierarchical principle

- Optimisation is based on the 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

- In practical terms:
  - we eliminate a variable if this ***reduces*** AIC
  - we test variables according to the hierarchical principle (higher-interactions first, single terms last)

The lowest AIC, the better the model

Example

```
> 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
```

- If we run a model of menarche from age and igf1 with interactions, neither *igf1* or *age:igf1* is significant

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 ***
```

- We must optimise model with function *step*

** we obtain more output than this; see code

Exercise

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

cat 842.31