Logistic Regression

Dr Tra My Pham
Department of Behavioural Science & Health
University College London

Email: tra.pham.09@ucl.ac.uk

Objectives

By the end of this session you should be able to:

- Use a logistic regression model
- Interpret the results of logistic regression
- Compare models using the likelihood ratio test
- Use interaction terms in logistic regression

Dependent variable

- It is a dichotomous variable
- It takes only two values, which usually represent the occurrence or non-occurrence of some outcome event (coded as 0 or 1)
- For example

    - CVD (0 “No” 1 “Yes”)

    - Mortality (0 “Alive” 1 “Dead”)

    - HIV (0 “Uninfected” 1 “Infected”)

Logistic regression

- It is a variation of ordinary (linear) regression
- The logistic regression model is used to explain the effects of the explanatory variable(s) on the binary response.
- The dependent variable (Y) is a dichotomous or binary variable
- The independent/explanatory variable(s) (X) can be continuous or binary/categorical

Logistic Regression

- Logistic regression models the log odds of having the event of interest

$$\ln(odds) = ln\left\{\frac{\hat{p}}{1-\hat{p}}\right\} = \beta_0 + x\beta_1,$$

where

- $\hat{p}$ is the observed probability of having the event
- β_0 is the intercept
- β_1 is the slope parameter

Logistic Regression

- We fit a regression model for the log odds of disease as the outcome measure
- The log odds can take any value, pos or neg, whereas risks (and probabilities) are constrained to lie between 0 and 1
- The model is fitted using the method of “Maximum likelihood” which is an iterative procedure

An example: a model with one independent variable

```
. logit cvddef1 sex

Iteration 0:   log likelihood = -6106.9672
Iteration 1:   log likelihood = -6103.9477
Iteration 2:   log likelihood = -6103.9464
Iteration 3:   log likelihood = -6103.9464

Logistic regression                             Number of obs     =     14,836
                                                LR chi2(1)        =       6.04
                                                Prob > chi2       =     0.0140
Log likelihood = -6103.9464                     Pseudo R2         =     0.0005
```

cvddef1	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	-.1154694	.0469308	-2.46	0.014	-.207452	-.0234867
_cons	-1.721946	.0343153	-50.18	0.000	-1.789202	-1.654689

- Women compared to men are less likely to have CVD
- The constant is the log odds of CVD when sex=0, i.e. for men

Odds ratio

- If the OR = 1 there is no association
- $0 \leq OR < 1$ means lower risk of disease
- $OR > 1$ means higher risk of disease

How to obtain the odds ratio

STATA COMMAND: **`logistic cvd sex`**

```
. logistic cvddef1 sex

Logistic regression                               Number of obs     =     14,836
                                                  LR chi2(1)        =       6.04
                                                  Prob > chi2       =     0.0140
Log likelihood = -6103.9464                       Pseudo R2         =     0.0005
```

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	.8909479	.0418129	-2.46	0.014	.8126522	.976787
_cons	.1787181	.0061328	-50.18	0.000	.1670934	.1911515

- Women compared to men are less likely to have CVD, exp(-0.115) = 0.89

How to interpret the results

- Among women the odds of having CVD is 0.89 times lower than men
- The p-value < 0.05
- The confidence interval tells us that there is a 95% chance that the interval [0.81, 0.97] captures the true OR

Testing for association

- We use the Wald test to test the null hypothesis that the true parameter value is 0 (i.e., in this case OR = 1, meaning that there is no association)
- z statistic is calculated as

$$z = \textbf{coefficient / SE}$$

$$z = \textbf{ln(OR) / SE (lnOR)}$$

we compare z with a Normal distribution

- For our example

$$z = -0.15 / 0.046 = -2.46$$

$p < 0.05$ we reject the null hypothesis of no association

Another example

- Dependent variable: CVD
- Independent variable: diabetes

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	diabetes	1.356	0.087	242.309	1	0.000	3.882	3.272	4.604
	Constant	-1.868	0.025	5751.755	1	0.000	0.154		

a. Variable(s) entered on step 1: diabetes.

How do we interpret the results?

A model with more than one independent variable

- Dependent variable: CVD
- Independent variables: sex and diabetes

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	diabetes	1.350	0.087	239.566	1	0.000	3.856	3.250	4.575
	sex	-0.094	0.047	3.917	1	0.048	0.910	0.830	0.999
	Constant	-1.816	0.036	2608.265	1	0.000	0.163		

a. Variable(s) entered on step 1: diabetes, sex.

How do we interpret the results?

Interpretation

- The logistic regression has produced simultaneously a summary estimate of the effect of diabetes adjusted for sex, and a summary estimate of the effect of sex adjusted for diabetes.
- 3.8=exp(1.35) represents the (summary) odds ratio of having CVD among those who have diabetes compared to those who don't, adjusted for any confounding effect of sex

An example with a continuous independent variable

m1 <- glm(cvddef1 ~ diabetes + sex + age, data = logit, family = binomial(link = "logit"))

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.243353   0.090688 -46.791  < 2e-16 ***
diabetes     0.832845   0.091848   9.068  < 2e-16 ***
sex         -0.161172   0.049628  -3.248  0.00116 **
age          0.046396   0.001451  31.972  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
                  odds      2.5 %     97.5 %
(Intercept) 0.01435937 0.01200256 0.01712679
diabetes    2.29985246 1.91897256 2.75103357
sex         0.85114554 0.77227223 0.93813868
age         1.04748865 1.04452898 1.05048824
```

- Note: Age is continuous
- The odds of having CVD is increasing by 1.05 times per each year increase in age (adjusted for sex and diabetes)

We can also add a categorical independent variable, cigarette smoking

m1 <- glm(cvddef1 ~ diabetes + sex + age + factor(cigst1), data = logit, family = binomial(link = "logit"))

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -4.294709   0.098359 -43.663  < 2e-16 ***
diabetes         0.818697   0.092065   8.893  < 2e-16 ***
sex             -0.111087   0.050659  -2.193   0.0283 *  
age              0.045027   0.001512  29.771  < 2e-16 ***
factor(cigst1)2 -0.019517   0.110426  -0.177   0.8597    
factor(cigst1)3  0.268925   0.059305   4.535 5.77e-06 ***
factor(cigst1)4  0.082462   0.068856   1.198   0.2311    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
                      odds      2.5 %     97.5 %
(Intercept)     0.01364054 0.01122987 0.01651344
diabetes        2.26754279 1.89121064 2.71353762
sex             0.89486092 0.81031304 0.98833822
age             1.04605566 1.04297419 1.04917671
factor(cigst1)2 0.98067184 0.78691380 1.21353669
factor(cigst1)3 1.30855733 1.16486777 1.46977530
factor(cigst1)4 1.08595749 0.94828028 1.24218780
```

Likelihood ratio test (LRT)

- $LRT = 2(L_1 - L_0)$
- L_1 is the log likelihood of the model with variable that you want to test
- L_0 is the log likelihood of the model without that variable
- Under the null hypothesis LRT is distributed as a Chi-square with 1 d.f. (because there is only one predictor)

Hypothesis testing

- Suppose we want to test the null hypothesis:

  H_0: after taking into account the effect of sex diabetes and age, there is no association between smoking and CVD

- We can use the LR test

In Stata…

- Obtain L_1 by fitting the model with smoking
```
logistic cvd diabetes sex age i.cist1
```

- Save L_1
```
estimates store a
```

- Obtain the value L_0 by fitting the model without smoking
```
logistic cvd diabetes sex age
```

- Save L_0
```
estimates store b
```

- Compare L_1 and L_0
```
lrtest a b
```

LRT result

```
lrtest a b
```

BUT!

The number of observations differs between models `a` (with smoking) and `b` (without smoking):

14764 vs. 14836

We need to re-run the model without smoking by excluding the people that have not answered to the question on smoking

```
. xi:logistic cvd diabetes sex age if cigst!=.

Logistic regression                             Number of obs     =     14,764
                                                LR chi2(3)        =    1362.84
                                                Prob > chi2       =     0.0000
Log likelihood = -5407.2266                     Pseudo R2         =     0.1119

------------------------------------------------------------------------------
     cvddef1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    diabetes |   2.298955   .2111569     9.06   0.000     1.920209    2.752406
         sex |   .8548839   .0424689    -3.16   0.002     .7755703    .9423085
         age |   1.047421   .0015244    31.83   0.000     1.044438    1.050413
       _cons |    .014396   .0013095   -46.62   0.000     .0120452    .0172056
------------------------------------------------------------------------------

. est store x

. lrtest a x

Likelihood-ratio test                                 LR chi2(3)  =     22.32
(Assumption: x nested in a)                           Prob > chi2 =    0.0001
```

Results

- $p < 0.05$ so we reject H_0
- After taking into account the effect of sex diabetes and age, there is an association between smoking and CVD

INTERACTIONS
(EFFECT MODIFICATION)

Analysis with two independent variables: CVD vs diabetes (yes/no) and age (old/young)

(d) had cardiovascular condition (excluding diabetes/high bp)	agegr <=50	51+	Total
no	7,706 93.39	4,998 75.90	12,704 85.63
yes	545 6.61	1,587 24.10	2,132 14.37
Total	8,251 100.00	6,585 100.00	14,836 100.00

Pearson chi2(**1**) = **910.9149** Pr = **0.000**

(d) had cardiovascular condition (excluding diabetes/high bp)	(d) doctor diagnosed diabetes (excluding pregnant) no	yes	Total
no	12,322 86.62	382 62.52	12,704 85.63
yes	1,903 13.38	229 37.48	2,132 14.37
Total	14,225 100.00	611 100.00	14,836 100.00

Pearson chi2(**1**) = **276.5524** Pr = **0.000**

Analysis with two independent variables: CVD vs diabetes (yes/no) and age (old/young)

Note: **agegr** is coded **0** for age ≤ 50 and **1** for age > 50

```
. logistic cvd diabetes agegr

Logistic regression                              Number of obs   =      14836
                                                 LR chi2(2)      =    1027.46
                                                 Prob > chi2     =     0.0000
Log likelihood =  -5593.237                      Pseudo R2       =     0.0841
```

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
diabetes	2.568399	.2314766	10.47	0.000	2.152524	3.064622
agegr	4.208948	.22483	26.91	0.000	3.790572	4.673501
_cons	.0693111	.0030804	-60.06	0.000	.0635291	.0756193

Interaction term

- In the previous model we estimated the joint effects of diabetes and age, assuming constant odds ratios across strata.
- If the odds ratios differ across strata then there is interaction between the two variables and the odds of the exposure should be reported separately for different levels of the effect modifying/interacting variable.

We can check by stratifying the analysis

- **xi:logistic cvd i.diabetes if agegr==0**

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Idiabetes_1	3.14372	.7698466	4.68	0.000	1.945352	5.080301
_cons	.0688658	.0031103	-59.24	0.000	.0630318	.0752398

- **xi:logistic cvd i.diabetes if agegr==1**

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Idiabetes_1	2.494317	.2399451	9.50	0.000	2.065708	3.011858
_cons	.292595	.0089581	-40.14	0.000	.2755538	.31069

- **xi:logistic cvd i.agegr if diabetes==0**

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Iagegr_1	4.248769	.2318272	26.51	0.000	3.817848	4.728328
_cons	.0688658	.0031103	-59.24	0.000	.0630318	.0752398

- **xi:logistic cvd i.agegr if diabetes==1**

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Iagegr_1	3.371091	.8676534	4.72	0.000	2.035578	5.582815
_cons	.216495	.0521067	-6.36	0.000	.135076	.3469906

Interaction in logistic regression (logit)

STATA COMMAND: `xi:logit cvd i.diabetes*i.agegrp`

```
Log likelihood = -5592.8637                          Pseudo R2       =     0.0842
```

cvddef1	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Idiabetes_1	1.145407	.2448839	4.68	0.000	.6654431	1.62537
_Iagegr_1	1.446629	.0545634	26.51	0.000	1.339687	1.553572
_IdiaXage_1_1	-.2313918	.2631007	-0.88	0.379	-.7470596	.284276
_cons	-2.675595	.0451644	-59.24	0.000	-2.764116	-2.587075

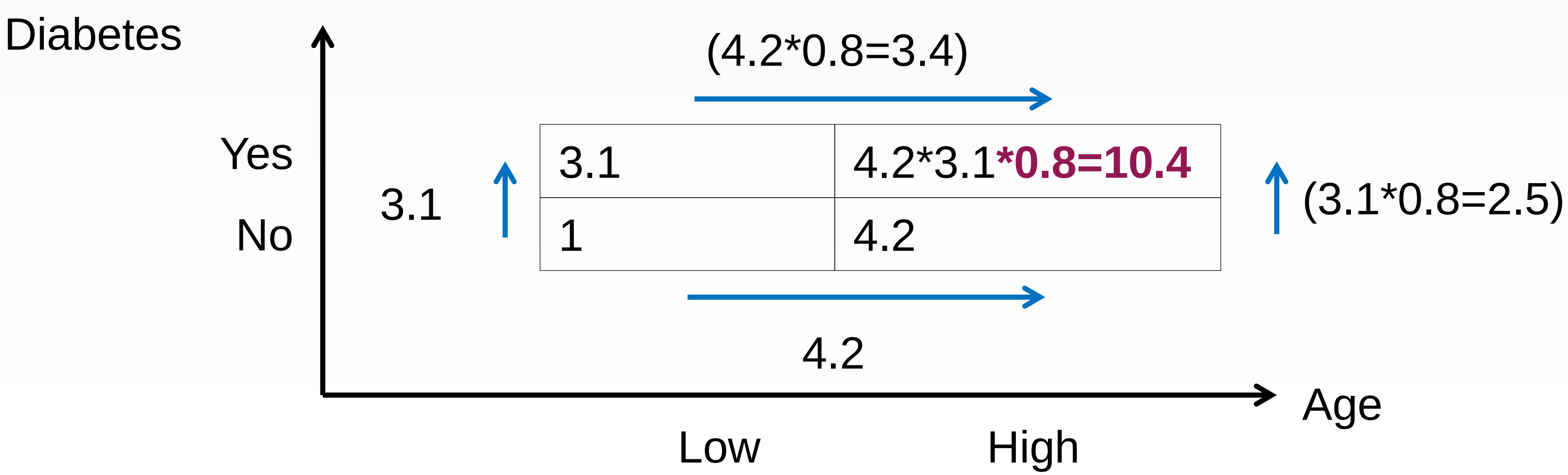
$$\ln(odds) = -2.7 + 1.1 * (diabetes = 1) + 1.4 * (agegrp = 1) + \underbrace{(-0.2) * (diabetes = 1) * (agegrp = 1)}_{\text{Interaction between } \mathbf{diabetes} \text{ and } \mathbf{agegrp}}$$

Interaction in logistic regression (logistic)

STATA COMMAND: `xi:logistic cvd i.diabetes*i.agegr`

```
Logistic regression                               Number of obs   =      14836
                                                  LR chi2(3)      =    1028.21
                                                  Prob > chi2     =     0.0000
Log likelihood = -5592.8637                       Pseudo R2       =     0.0842

------------------------------------------------------------------------------
     cvddef1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Idiabetes_1 |   3.14372    .7698465     4.68   0.000     1.945352    5.080301
    _Iagegr_1 |  4.248769    .2318272    26.51   0.000     3.817848    4.728327
 _IdiaXage_1_1 |  .7934285    .2087516    -0.88   0.379     .4737575      1.3288
        _cons |  .0688658    .0031103   -59.24   0.000     .0630318    .0752398
------------------------------------------------------------------------------
```



Interpretation

- The interaction term between diabetes and age has odds ratio 0.8
  - The odds ratio of age is different in people with and without diabetes
  - The odds ratio of diabetes is different between younger and older
- Among those aged ≤ 50 (the baseline of age) the odds ratio for diabetes vs no-diabetes is 3.1
- Among those aged 51+ (not at the baseline) the odds ratio for diabetes vs no-diabetes is 3.1 multiplied by the interaction parameter 0.8 = 2.5
- Among no-diabetes (the baseline of diabetes), the odds ratio for high age (51+ vs ≤50) is 4.2
- Among diabetes, the odds ratio for high age (51+ vs ≤ 50) is 4.2 multiplied by 0.8 = 3.4

Important

In the model **without** interaction between diabetes and age, the parameter

OR for diabetes = 2.6

is interpreted as the summary odds ratio for diabetes adjusted for the effect of age

In the model **with** interaction term between diabetes and age, the parameters

OR for diabetes = 3.1 in the agegr = 0 stratum

OR for diabetes = 2.5 in the agegr = 1 stratum

are interpreted as a stratum specific odds ratios: the odds ratios for diabetes in each stratum of age

Another example with physical activity (low=0, high=1)

(d) had cardiovascular condition (excluding diabetes/high bp)	pact 0	1	Total
no	8,746 83.14	3,915 91.64	12,661 85.60
yes	1,773 16.86	357 8.36	2,130 14.40
Total	10,519 100.00	4,272 100.00	14,791 100.00

(d) had cardiovascular condition (excluding diabetes/high bp)	sex men	women	Total
no	5,601 84.84	7,103 86.26	12,704 85.63
yes	1,001 15.16	1,131 13.74	2,132 14.37
Total	6,602 100.00	8,234 100.00	14,836 100.00

```
Logistic regression                             Number of obs     =     14,791
                                                LR chi2(2)        =     211.85
                                                Prob > chi2       =     0.0000
Log likelihood =  -5990.484                     Pseudo R2         =     0.0174
```

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	.8215553	.0390729	-4.13	0.000	.7484346	.9018196
pact	.4378771	.0269326	-13.43	0.000	.3881479	.4939777
_cons	.2270913	.0084758	-39.72	0.000	.2110721	.2443263

logistic cvd sex if pact==0

.

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	.7683977	.0402843	-5.03	0.000	.693363	.8515525
_cons	.2357019	.0091722	-37.14	0.000	.2183931	.2543825

logistic cvd sex if pact==1

.

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	1.10931	.1227473	0.94	0.348	.8930298	1.377971
_cons	.0868334	.0066736	-31.80	0.000	.0746909	.10095

logistic cvd pact if sex==0

.

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
pact	.3684035	.0317363	-11.59	0.000	.3111692	.4361651
_cons	.2357019	.0091722	-37.14	0.000	.2183931	.2543825

logistic cvd pact if sex==1

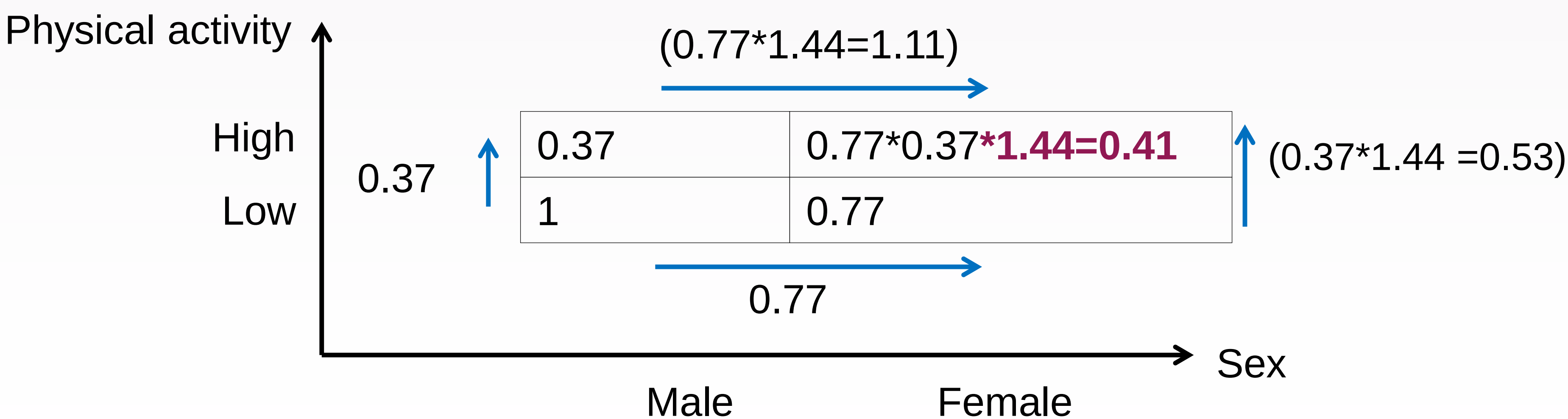
.

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
pact	.5318518	.0462782	-7.26	0.000	.4484611	.630749
_cons	.1811128	.0063627	-48.64	0.000	.1690619	.1940227

Logistic regression with interaction

<u>STATA COMMAND:</u> `xi:logistic cvd i.sex*i.pact`

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Isex_1	.7683977	.0402843	-5.03	0.000	.693363	.8515525
_Ipact_1	.3684035	.0317363	-11.59	0.000	.3111692	.4361651
_IsexXpac_1_1	1.443666	.1767673	3.00	0.003	1.135646	1.835231
_cons	.2357019	.0091722	-37.14	0.000	.2183931	.2543825



Interpretation

- The interaction term between sex and pact has odds ratio 1.44
  - The odds ratio for physical activity is different between men and women
  - The odds ratio fir sex is different between active and non-active

- Among those less active (the baseline of pact) the odds ratio for women vs men is 0.77
- Among those active (not at the baseline) the odds ratio for women vs men is 0.77 multiplied by the interaction parameter 1.44 = 1.11

- Among men (the baseline of sex), the odds ratio for the effect of pact (active vs less active) is 0.37
- Among women, the odds ratio for the effect of pact (active vs less active) is 0.37 multiplied by 1.44 = 0.53

Again, this can be confirmed by stratified analyses

Note!

The odds ratio for sex diverges according to the different levels of physical activity

- 0.77 (< 1) in less active
- 1.1 (> 1) in highly active

but for both genders, the modifiable risk factor of high (vs low) activity is beneficial

Important

- In the first example of interaction between diabetes and age, the baseline ORs are greater than 1 and the interaction is negative (OR < 1)
  - this implies that the effect of diabetes tends to decrease with age (3.1 and 2.5)

- In the second example of interaction between sex and physical activity, the baseline ORs are less than 1 and the interaction is positive (OR > 1)
  - this implies that the effect of sex is diverging according to different levels of physical activity (0.77 and 1.1)

Interaction with a continuous variable

m1 <- glm(cvddef1 ~ diabetes + age + diabetes:age, data = logit, family = binomial(link = "logit"))

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)  -4.361723   0.090346 -48.278  < 2e-16 ***
diabetes      1.805679   0.421565   4.283 1.84e-05 ***
age           0.046899   0.001487  31.538  < 2e-16 ***
diabetes:age -0.014803   0.006401  -2.313   0.0207 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
                   odds      2.5 %      97.5 %
(Intercept)  0.01275639 0.01066833  0.01520261
diabetes     6.08410154 2.60797383 13.64613360
age          1.04801623 1.04498230  1.05109221
diabetes:age 0.98530604 0.97323994  0.99800089
```

Interpretation

- Diabetes = 6.08 is the odds ratio for CVD in those with diabetes vs no-diabetes, holding age constant
- Age = 1.04 is the increase in the odds of CVD per each year increase in age, amongst non-diabetics
- diabetes:age = 0.98 is the interaction term between diabetes and age (one year increase)

- Note for STATA users: using **`i.diabetes*age`** tells STATA we want interaction between **diabetes** and **age** (where **age** is continuous)
  - Beware: STATA doesn't understand `age*i.diabetes`, only **`i.diabetes*age`**

LIKELIHOOD RATIO TEST FOR INTERACTIONS

Likelihood ratio test

- We can perform the LR test for the null hypothesis that there is no interaction
- We do this by comparing the log likelihoods of the model with the interaction term and the model without
  - LRT = 2 (L_1 - L_0)
  - L_1 is the log likelihood of model with the variable that you want to test
  - L_0 is the log likelihood of the model without that variable

Interaction term involving a variable with more than 2 categories

```
. xi:logistic cvd i.sex*i.smoking
i.sex             _Isex_0-1         (naturally coded; _Isex_0 omitted)
i.smoking         _Ismoking_0-3     (naturally coded; _Ismoking_0 omitted)
i.sex*i.smoking   _IsexXsmo_#_#     (coded as above)

Logistic regression                            Number of obs     =     14,764
                                               LR chi2(5)        =     232.55
                                               Prob > chi2       =     0.0000
Log likelihood = -5972.3701                    Pseudo R2         =     0.0191
```

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Isex_1	1.092743	.0797532	1.22	0.224	.9470947	1.260789
_Ismoking_2	2.346264	.1824539	10.97	0.000	2.01458	2.732558
_Ismoking_3	.8582719	.08505	-1.54	0.123	.7067658	1.042256
_IsexXsmo_1_2	.7069234	.0766261	-3.20	0.001	.5716199	.8742534
_IsexXsmo_1_3	.9895677	.1287282	-0.08	0.936	.7668611	1.276951
_cons	.1342282	.0077527	-34.77	0.000	.1198616	.1503167

```
. est store a
```

LRT test

Null hypothesis: There is no interaction between sex and smoking status

We test it with LRT, if $p < 0.05$ we reject the null hypothesis

```
. xi:logistic cvd i.sex i.smoking
i.sex             _Isex_0-1           (naturally coded; _Isex_0 omitted)
i.smoking         _Ismoking_0-3       (naturally coded; _Ismoking_0 omitted)

Logistic regression                              Number of obs     =     14,764
                                                 LR chi2(3)        =     220.75
                                                 Prob > chi2       =     0.0000
Log likelihood = -5978.2724                      Pseudo R2         =     0.0181
```

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Isex_1	.9616179	.0459716	-0.82	0.413	.8756077	1.056077
_Ismoking_2	1.970778	.105413	12.68	0.000	1.774633	2.188602
_Ismoking_3	.8460627	.0542723	-2.61	0.009	.7461061	.9594106
_cons	.1452046	.0066206	-42.32	0.000	.1327913	.1587782

```
. est store b

. lrtest a b

Likelihood-ratio test                                 LR chi2(2)  =     11.80
(Assumption: b nested in a)                           Prob > chi2 =    0.0027
```

$p < 0.05$ => we reject the null hypothesis of no interaction

. xi:logistic cvd i.smoking if sex==0

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Ismoking_2	2.346265	.182454	10.97	0.000	2.014581	2.732559
_Ismoking_3	.8582718	.08505	-1.54	0.123	.7067658	1.042256
_cons	.1342282	.0077527	-34.77	0.000	.1198616	.1503167

. xi:logistic cvd i.smoking if sex==1

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Ismoking_2	1.65863	.1252463	6.70	0.000	1.430453	1.923204
_Ismoking_3	.8493182	.0715772	-1.94	0.053	.7200032	1.001858
_cons	.1466769	.0065444	-43.02	0.000	.1343949	.1600812

. xi:logistic cvd sex if smoking==0

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	1.092742	.0797531	1.22	0.224	.9470943	1.260788
_cons	.1342283	.0077527	-34.77	0.000	.1198617	.1503168

. xi:logistic cvd sex if smoking==1

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	.7724846	.0619073	-3.22	0.001	.6601979	.9038691
_cons	.3149351	.0163984	-22.19	0.000	.2843804	.3487727

. xi:logistic cvd sex if smoking==2

cvddef1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	1.081343	.1164413	0.73	0.468	.8755966	1.335434
_cons	.1152043	.0092764	-26.84	0.000	.0983849	.1348991

Interpretation

- `_Isex_1:` the odds of having CVD is 1.09 times higher in women compared to men in the non-smokers category

- `_Ismoking_2:` the odds of having CVD is 2.34 times higher among men who are ex-smokers compared to men who are non-smokers. For women, this is 2.34 * 0.71 = 1.65

- `_Ismoking_3:` the odds of having CVD is 0.85 times lower among men who are current smokers compared to men who are non-smokers For women, this is 0.86 * 0.99 = 0.85

- `_IsexXsmo_1_2 and _IsexXsmo_1_3` are interaction terms

- OR comparing women to men among ex-smokers 1.09*0.71 = 0.77. For current smokers, this is 1.09*0.99 = 1.08

Suggested readings

- Tabachnick B, Fidell L. Using Multivariate Statistics. 4th Edition. London, Allyn & Bacon, 2001

- Clayton D, Hills M. Statistical Methods in Epidemiology. Oxford University Press 1993

- Hosmer D W, Lemeshow S. Applied logistic regression. 2nd Edition. New York, Chichester Wiley, 2000

- Long JS, Freese J Regression Models for Categorical Dependent Variables Using Stata, 2nd edition, Stata Press, 2006