Introduction to logistic regression

Tuan V. Nguyen
Professor and NHMRC Senior Research Fellow
Garvan Institute of Medical Research
University of New South Wales
Sydney, Australia

What we are going to learn

- Uses of logistic regression model
- Probability, odds, logit
- Estimation and interpretation of parameters

Consider a case-control study

	Lung Cancer	Controls
Smokers	647	622
Non-smokers	2	27

R Doll and B Hill. BMJ 1950; ii:739-748

- How can we show the association between smoking and lung cancer risk?

Risk factors for fracture: prospective study

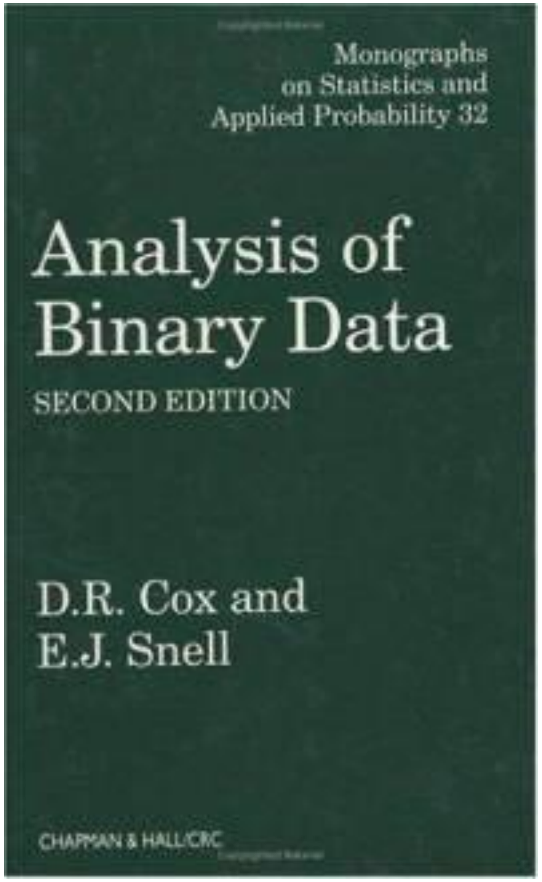
id	sex	fx	durfx	age	wt	ht	bmi	Tscores	fnbmd	lsbmd	fall	priorfx	death
3	M	0	0.55	73	98	175	32	0.33	1.08	1.458	1	0	1
8	F	0	15.38	68	72	166	26	-0.25	0.97	1.325	0	0	0
9	M	0	5.06	68	87	184	26	-0.25	1.01	1.494	0	0	1
10	F	0	14.25	62	72	173	24	-1.33	0.84	1.214	0	0	0
23	M	0	15.07	61	72	173	24	-1.92	0.81	1.144	0	0	0
24	F	0	12.3	76	57	156	23	-2.17	0.74	0.98	1	0	1
26	M	0	11.47	63	97	173	32	-0.25	1.01	1.376	1	0	1
27	F	0	15.13	64	85	167	30	-1.17	0.86	1.073	0	0	0
28	F	0	15.08	76	48	153	21	-2.92	0.65	0.874	0	0	0
29	F	0	14.72	64	89	166	32	-0.17	0.98	1.088	0	0	0
32	F	0	14.92	60	105	165	39	-0.33	0.96	1.154	3	0	0
33	F	0	14.67	75	52	156	21	-1.42	0.83	0.852	0	0	0
34	F	1	1.64	75	70	160	27	-1.75	0.79	1.186	0	0	0
36	M	0	15.32	62	97	171	33	1	1.16	1.441	0	0	0
37	F	0	15.32	60	60	161	23	-1.75	0.79	0.909	0	0	0

- **Dubbo Osteoporosis Epidemiology Study**
- **Question: what are predictors of *fracture risk***

Uses of logistic regression

- To describe relationships between outcome (dependent variable) and risk factors (independent variables)
- Controlling for confounders
- Developing prognostic models

Logistic regression model



1970

Professor David R. Cox
Imperial College, London

Some examples of logistic regression

Identification of undiagnosed type 2 diabetes by systolic blood pressure and waist-to-hip ratio

M. T. T. Ta · K. T. Nguyen · N. D. Nguyen · L. V. Campbell · T. V. Nguyen

Table 2 Association between risk factor and type 2 diabetes: univariate logistic regression analysis

Risk factor	Comparison unit[a]	Men		Women	
		OR (95% CI)	c statistic	OR (95% CI)	c statistic
Age (years)	5	1.28 (1.05–1.56)	0.58	1.19 (1.05–1.36)	0.56
Weight (kg)	10	1.57 (1.26–1.96)	0.64	1.53 (1.30–1.81)	0.61
Waist circumference (cm)	10	1.89 (1.48–2.40)	0.69	1.60 (1.37–1.86)	0.63
WHR	0.07	2.54 (1.85–3.50)	0.71	1.72 (1.46–2.03)	0.64
Lean mass (kg)	7	1.46 (1.08–1.96)	0.59	1.36 (1.00–1.85)	0.55
Fat mass (kg)	7	1.84 (1.43–2.38)	0.66	1.60 (1.36–1.88)	0.62
Per cent body fat	10	2.29 (1.61–3.28)	0.66	2.01 (1.54–2.65)	0.62
Abdominal fat (kg)	4	1.77 (1.38–2.27)	0.65	1.58 (1.35–1.84)	0.63
Systolic BP (mmHg)	20	1.62 (1.32–2.00)	0.65	1.50 (1.31–1.73)	0.63
Diastolic BP (mmHg)	12	1.44 (1.16–1.79)	0.62	1.40 (1.21–1.61)	0.61

[a]The comparison unit was set to be close to the standard deviation of each risk factor

Some examples of logistic regression

- "This study identified behavioral and psychosocial/ interpersonal factors in young adolescence that are associated with handgun carrying in later adolescence."

TABLE 3—Logistic Regression Analysis of Behavioral Variables Measured in 9th Grade Predicting Handgun Carrying in 12th Grade among Students in San Diego and Los Angeles Counties

	No.	Boys, Odds Ratio (95% CI)	Girls, Odds Ratio (95% CI)
Days absent from school in previous month (unrelated to illness)			
0	1235	1.00	1.00
1–2	462	1.40 (0.92, 2.13)	0.78 (0.41, 1.51)
3 or more	243	2.37 (1.50, 3.73)	0.91 (0.39, 2.11)
Grades			
Mostly A's or A's and B's	986	1.00	1.00
Mostly B's or B's and C's	804	0.95 (0.65, 1.36)	1.74 (0.96, 3.15)
Mostly C's or below	399	1.31 (0.87, 1.98)	1.97 (0.97, 4.00)

When to use logistic regression?

- **Logistic regression**:
  - outcome is a categorical variable (usually binary – yes/no)
  - risk factors are either continuous or categorical variables

- **Linear regression**:
  - outcome is a continuous variable
  - risk factors are either continuous or categorical variables

Logistic regression and Odds

- Linear regression works on continuous data
- Logistic regression works on odds of an outcome

Risk, probability and odds

- Risk: probability (P) of an event [during a period]
- Odds: ratio of probability of having an event to the probability of not having the event

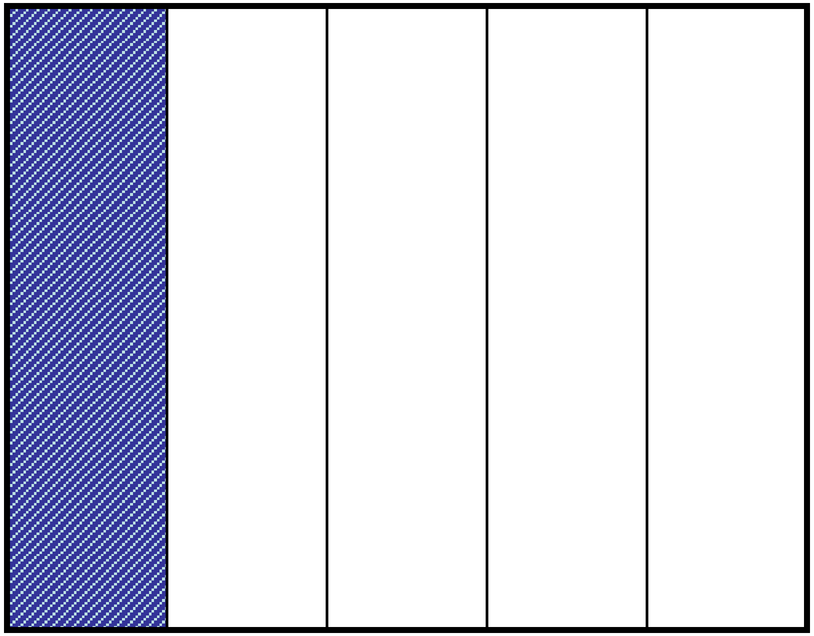
$$\text{Odds} = P \,/\, (1 - P)$$

- One out of 5 patients suffer a stroke …

$P = 1/ 5 = 0.20$

$\text{Odds} = 0.2 / 0.8 = 1 \text{ to } 4$

Probability and odds



- P = 1/5 = 0.2 or 20%
- Odds = (P) / (1-P)
- Odds = 0.2 / 0.8 or 1:4 or "one to four"

Probability, odds, and logit

- **Probability**: from 0 to 1
- **Odds**: continuous variable
  - When Probability = 0.5, odds = 1
- **Logit = log odds**

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

The logistic regression model

- **Let X be a risk factor**
- **Let P be the probability of an event (outcome)**
- **The logistic regression model is defined as:**

$$\mathrm{logit}(\mathrm{p}) = a + bX$$

or

$$\log\left(\frac{p}{1-p}\right) = a + bX$$

The logistic regression model

$$\log\left(\frac{p}{1-p}\right) = a + bX$$

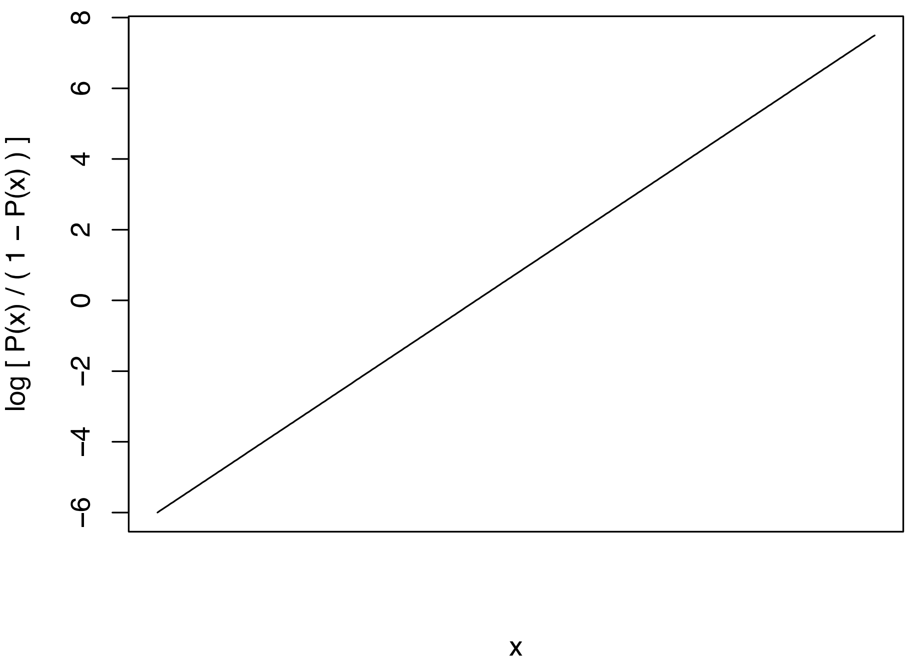
That also means:

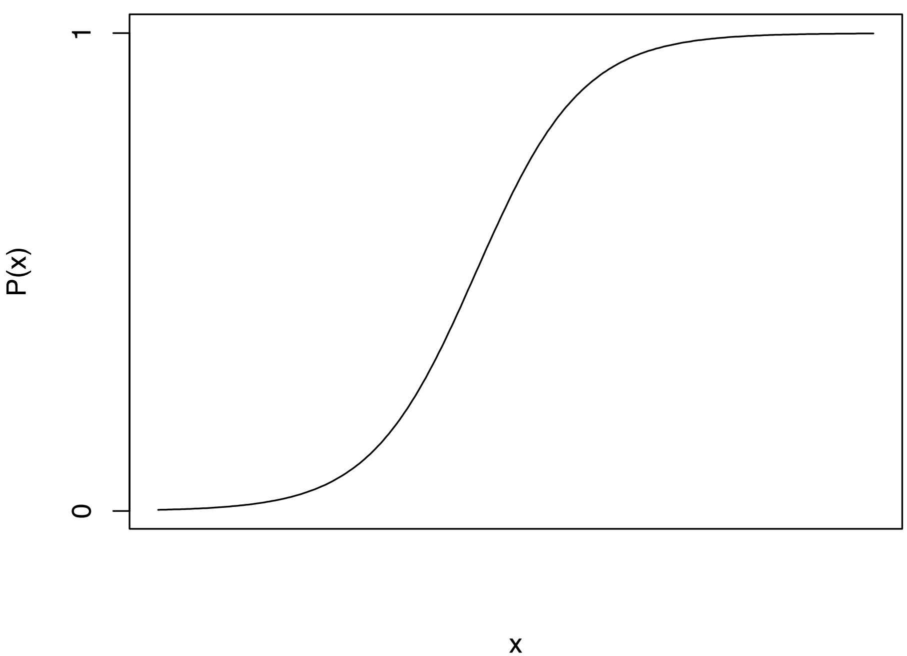
$$p = \frac{e^{a+bX}}{1+e^{a+bX}}$$

Relationship between X, p and logit(p)

$$\log\left(\frac{p}{1-p}\right) = a + bX$$

$$p = \frac{e^{a+bX}}{1+e^{a+bX}}$$





Meaning of logistic regression parameters

$$\log\left(\frac{p}{1-p}\right) = a + bX$$

- α **is the log odds of the outcome for X = 0**
- β **is the log odds ratio associated with a unit increase in X**

- **Odds ratio = exp(β)**

Assumptions of logistic regression model

- Model provides an appropriate representation for the dependence of outcome probability on predictor(s)
- Outcomes are independent
- Predictors measured without error

Advantages of logistic regression model

- Outcome probability changes smoothly with increasing values of predictor, valid for arbitrary predictor values
- Coefficients are interpreted as log odds ratios
- Can be applied to a range of study designs (including case- control)
- Software widely available

Analysis of case control study

Consider a case-control study

	Lung Cancer	Controls
Smokers	647	622
Non-smokers	2	27

R Doll and B Hill. BMJ 1950; ii:739-748

Manual calculation of odds ratio

	Disease	No disease
Risk +ve	*a*	*b*
Risk –ve	*c*	*d*

	Lung K	Control
Smoking	647	622
No smoking	2	27

$$OR = \frac{ad}{bc}$$

$$LOR = \log(OR)$$

$$SE(LOR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$$95\%CI(LOR) = LOR \mp 1.96SE(LOR)$$

$$95\%CI(OR) = e^{LOR \mp 1.96SE(LOR)}$$

$$OR = \frac{647 \times 27}{622 \times 2} = 14.04$$

$$LOR = \log(14.04) = 2.64$$

$$SE(LOR) = \sqrt{\frac{1}{647} + \frac{1}{622} + \frac{1}{2} + \frac{1}{27}} = 0.735$$

$$95\%CI(LOR) = 2.642 \mp 1.96 \times 0.735$$

$$95\%CI(OR) = e^{2.64 \mp 1.96 \times 0.735}$$

= 3.32 to 59.03

Analysis by logistic regression model

- **P = probability of cancer (0 = No cancer, 1 = Cancer)**
- **X = smoking status (0 = No, 1 = Yes)**
- **Logistic regression model**

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta X$$

- **We want to estimate α and β**

R codes

	Lung K	Control
Smoking	647	622
No smoking	2	27

```
noyes = c(1, 0) # define a variable with 2 values 1=yes, 0=no
smoking = gl(2,1, 4, noyes) # smoking
cancer =  gl(2,2, 4, noyes) # cancer
ntotal = c(647, 2, 622, 27)  # actual number of patients
res = glm(cancer ~ smoking, family=binomial, weight=ntotal)
summary(res)
```

R codes (longer way)

	Lung K	Control
Smoking	647	622
No smoking	2	27

```
cancer = c(1, 1, 0, 0)
smoking = c(1, 0, 1, 0)
ntotal = c(647, 2, 622, 27)  # actual number of patients
res = glm(cancer ~ smoking, family=binomial, weight=ntotal)
summary(res)
```

R codes (rms package)

	Lung K	Control
Smoking	647	622
No smoking	2	27

```
cancer = c(1, 1, 0, 0)
smoking = c(1, 0, 1, 0)
ntotal = c(647, 2, 622, 27)  # actual number of patients
res = lrm(cancer ~ smoking, weight=ntotal)
summary(res)
```

R results

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.6027     0.7320  -3.556 0.000377 ***
smoking       2.6421     0.7341   3.599 0.000319 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1799.4  on 3  degrees of freedom
Residual deviance: 1773.3  on 2  degrees of freedom
AIC: 1777.3
```

R results

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.6027     0.7320  -3.556 0.000377 ***
smoking       2.6421     0.7341   3.599 0.000319 ***
```

- The model is:

$$\log\left(\frac{p}{1-p}\right) = -2.60 + 2.64 \times smoking$$

- Note that the coefficient for smoking is 2.64 (exactly the same with manual calculation)
- That is log(odds ratio) = 2.64
- Odds ratio = exp(2.64) = 14.01

Calculating odds ratio (OR)

```
cancer = c(1, 1, 0, 0)

smoking = c(1, 0, 1, 0)

ntotal = c(647, 2, 622, 27)  # actual number of patients

res = glm(cancer ~ smoking, family=binomial,
weight=ntotal)


library(epicalc)
logistic.display(res)
```

Calculating odds ratio (OR) and 95% CI

```
> logistic.display(res)

Logistic regression predicting cancer

                  OR(95%CI)             P(Wald's test) P(LR-
test)
smoking: 1 vs 0  14.04 (3.33,59.2)  < 0.001          < 0.001

Log-likelihood = -886.6352
No. of observations = 4
AIC value = 1777.2704
```

Analysis of raw data

Formal description of logistic regression

- Let Y be a binary response variable
  - $Y_i = 1$ if the trait is present in observation (person, unit, etc...) i
  - $Y_i = 0$ if the trait is NOT present in observation i
- $X = (X_1, X_2, ..., X_k)$ be a set of explanatory variables which can be discrete, continuous, or a combination. xi is the observed value of the explanatory variables for observation i.

Formal description of logistic regression

- **The logistic regression model is:**

$$p_i = \Pr(Y_i = 1 \mid X_i = x_i) = \frac{\exp(b_0 + b_i x_i)}{1 + \exp(b_0 + b_i x_i)}$$

- **Or, in logit expression:**

$$\operatorname{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots$$

Assumptions of logistic regression

- The data $Y_1, Y_2, \ldots, Y_n$ are independently distributed
- Distribution of Y_i is $Bin(n_i, \pi_i)$, i.e., binary logistic regression model assumes binomial distribution of the response
- Linear relationship between the logit of the explanatory variables and the response; $logit(\pi) = \beta_0 + \beta X$.
- The homogeneity of variance does NOT need to be satisfied
- Errors need to be independent but NOT normally distributed

Assessment of goodness-of-fit

- Overall goodness-of-fit statistics of the model;
- Pearson chi-square statistic, χ^2
- Deviance, G^2
- Likelihood ratio test, and statistic, ΔG^2
- Hosmer-Lemeshow test and statistic
- Residual analysis: Pearson, deviance, adjusted residuals, etc
- Overdispersion

Parameter estimation

- The *maximum likelihood estimator* (MLE) for (β0, β1) is obtained by finding ( ) that maximizes

$$L\left(b_0, b_1\right) = \prod_{i=1}^{N} p_i^{y_i}\left(1 - p_i\right)^{n_i - y_i} = \prod_{i=1}^{N} \frac{\exp\left(y_i\left(b_0 + b_1 x_i\right)\right)}{1 + \exp\left(b_0 + b_1 x_i\right)}$$

- This is implemented in R program called “glm” and “lrm”

Function glm in R

- General format

```
res= glm(outcome ~ riskfactor, family=binomial)
```

- `outcome` has values (0, 1)
- `riskfactor` has any value
- To get odds ratio and 95% CI

```
library(epicalc)
logistic.display(res)
```

Function glm in R

- To get goodness of fit of a model, use rms package

```
library(rms)
res = lrm(outcome ~ riskfactor)
summary(res)
```

An example of analysis: fracture data

id	sex	fx	durfx	age	wt	ht	bmi	Tscores	fnbmd	lsbmd	fall	priorfx	death
3	M	0	0.55	73	98	175	32	0.33	1.08	1.458	1	0	1
8	F	0	15.38	68	72	166	26	-0.25	0.97	1.325	0	0	0
9	M	0	5.06	68	87	184	26	-0.25	1.01	1.494	0	0	1
10	F	0	14.25	62	72	173	24	-1.33	0.84	1.214	0	0	0
23	M	0	15.07	61	72	173	24	-1.92	0.81	1.144	0	0	0
24	F	0	12.3	76	57	156	23	-2.17	0.74	0.98	1	0	1
26	M	0	11.47	63	97	173	32	-0.25	1.01	1.376	1	0	1
27	F	0	15.13	64	85	167	30	-1.17	0.86	1.073	0	0	0
28	F	0	15.08	76	48	153	21	-2.92	0.65	0.874	0	0	0
29	F	0	14.72	64	89	166	32	-0.17	0.98	1.088	0	0	0
32	F	0	14.92	60	105	165	39	-0.33	0.96	1.154	3	0	0
33	F	0	14.67	75	52	156	21	-1.42	0.83	0.852	0	0	0
34	F	1	1.64	75	70	160	27	-1.75	0.79	1.186	0	0	0
36	M	0	15.32	62	97	171	33	1	1.16	1.441	0	0	0
37	F	0	15.32	60	60	161	23	-1.75	0.79	0.909	0	0	0

- **Filename: `fracture.csv`**
- **Question: what are effects of age, weight, sex on *fracture risk***

R analysis

```
setwd("/Users/tuannguyen/Documents/_Vietnam2012/Can Tho /Datasets") # can also use file.choose()
fract = read.csv("fracture.csv", na.string=".”, header=T)
attach(fract)
names(fract)

library(rms)
dat = datadist(fract)
options(datadist="dat")
res = lrm(fx ~ sex)
summary(res)
```

Effect of sex on fracture risk

```
> res = lrm(fx ~ sex)
> summary(res)

              Effects                   Response : fx

 Factor       Low High Diff. Effect S.E. Lower 0.95 Upper 0.95
 sex - M:F    1   2    NA    -0.78  0.11 -0.99      -0.57
  Odds Ratio  1   2    NA     0.46    NA  0.37       0.57
```

- Men had lower ODDS of fracture than women (OR 0.46; 95% CI: 0.37 to 0.57)

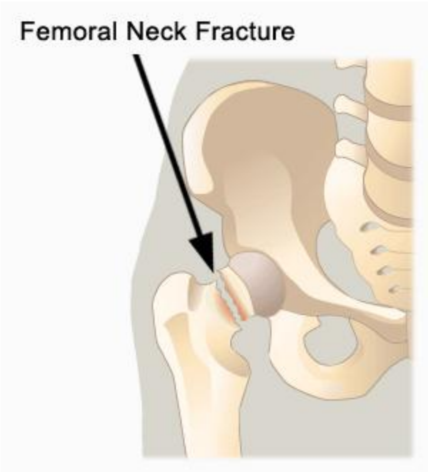
More on R output ...

```
> res


                     Model Likelihood     Discrimination    Rank Discrim.
                        Ratio Test            Indexes          Indexes
Obs            2216    LR chi2      55.76    R2       0.036    C       0.586
 0             1641    d.f.             1    g        0.369    Dxy     0.173
 1              575    Pr(> chi2) <0.0001    gr       1.446    gamma   0.370
max |deriv| 1e-11                            gp       0.066    tau-a   0.066
                                             Brier    0.187


          Coef    S.E.   Wald Z Pr(>|Z|)
Intercept -0.7829 0.0585 -13.39 <0.0001
sex=M     -0.7770 0.1074  -7.23 <0.0001
```

Effect of bone mineral density on fracture risk



- Bone mineral density measured at the femoral neck (`fnbmd`)
- Values: 0.28 to 1.51 g/cm^2
- Lower FNBMD increases the risk of fracture
- We want to estimate the odds ratio of fracture associated with FNBMD

R analysis

```
> res = lrm(fx ~ fnbmd)
> summary(res)
              Effects                    Response : fx


 Factor       Low  High Diff. Effect S.E. Lower 0.95 Upper 0.95
 fnbmd        0.73 0.93 0.2   -0.96  0.08 -1.11      -0.81
  Odds Ratio 0.73 0.93 0.2    0.38    NA  0.33       0.45
```

- **Each standard deviation increase in FNBMD is associated with a 72% reduction in the odds of fracture (OR 0.38; 95% CI 0.33 to 0.45)**

Summary

- Logistic regression model is very useful for
  - Decsribing relationship between an outcome and risk factors
  - Developing prognostic models in medicine
- Logistic regression model is applied when
  - Outcome is a categorical variable
- Logistic regression model is applicable to all study desgns, but mainly case control study