8. Logistic Regression

Introduction to Medical Statistics

OUCRU, Ho Chi Minh City

March 23-27, 2026

Nguyen Lam Vuong

and the biostatistics crew

Learning Objectives

- Know risk/odds, risk ratio (RR)/odds ratio (OR)
- Understand the logistic regression model
- Can interpret the results from univariable and multivariable logistic regression model

Risk and odds.
Three effect measures

Example

- Results from an influenza vaccine trial during an epidemic

	Influenza		Total
	Yes	No	
Vaccine	20 (8.3%)	220 (91.7%)	240
Placebo	80 (36.4%)	140 (63.6%)	220

Kirkwood & Sterne, p. 150

Probabilities (risks) and odds

	Influenza		Total
	Yes	No	
Vaccine	20 (8.3%)	220 (91.7%)	240
Placebo	80 (36.4%)	140 (63.6%)	220

- Probability/risk p
  - Estimated as (# influenza cases) / (# n total)
  - 8.3% (20/240) for vaccine
  - 36.4% (80/220) for placebo
- Odds
  - General definition: odds= p/(1-p)
  - Estimated as (# influenza cases)/(# non-influenza cases)
  - 0.09 = 1:11 ((20/240)/(220/240)=20/220) for vaccine
  - 0.57 = 8:14 (80/140) for placebo

Risk difference, relative risk, odds ratio

	Influenza		Total
	Yes	No	
Vaccine	20 (8.3%)	220 (91.7%)	240
Placebo	80 (36.4%)	140 (63.6%)	220

- Risk difference: $p_{vaccine} - p_{placebo}$: -0.281

  "risk of influenza 0.281 (28.1%) lower for vaccinated patients"

- Relative risk: $p_{vaccine} / p_{placebo}$: 0.229

  "vaccination leads to 1/0.229=4.4 fold reduction in risk of influenza"
  "risk of influenza for vaccinated patients 22.9% of risk for unvaccinated patients", but:
  "*risk of influenza 77.1% lower for vaccinated patients*" is ambiguous

- Odds ratio (OR): $odds_{vaccine} / odds_{placebo}$: 0.159

  "vaccination leads to 1/0.159=6.3 fold reduction in odds of influenza"

Interpreting the RR/OR

- OR/RR = 1: event equally likely in both groups
- OR/RR > 1: event more likely in first (numerator) group
- OR/RR < 1: event less likely in first (numerator) group

Use

- Risk difference
  - Easy to understand, translates directly to clinical decision making

    [If vaccine reduces absolute risk by 28%, then vaccinating 100 subjects will prevent ~28 influenzas.]

- Relative risk (RR)
  - Easy to understand, popular in epidemiology

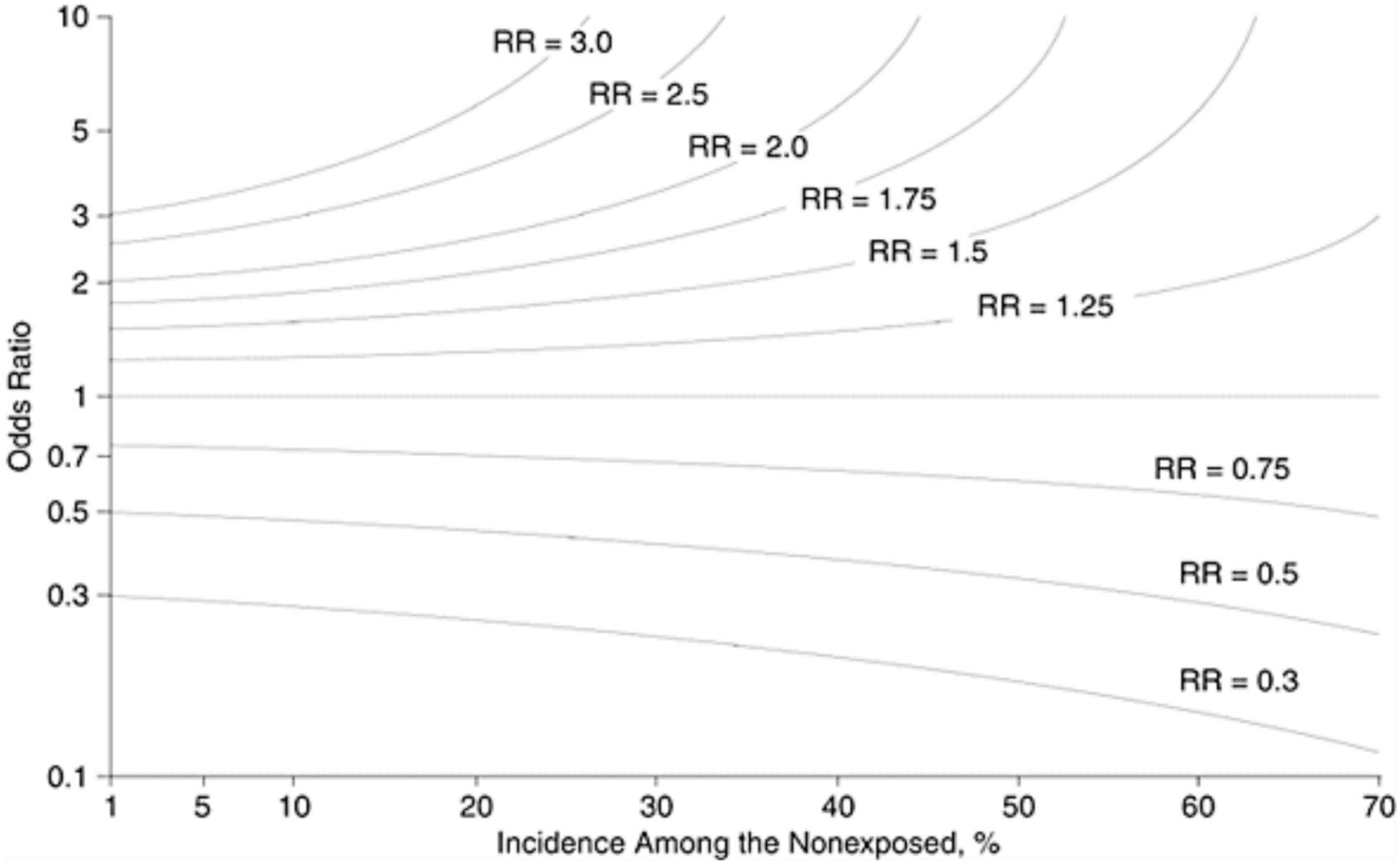
- Odds ratio (OR)
  - Mathematically the ‘natural choice’ and the target of logistic regression
  - Value further away from 1 than RR
  - Don’t interpret OR as RR, unless low event risk in population (say < 5%)

RR vs OR

Study	Risk/Probability		Odds		RR vs OR	
	Group 1 (p_1)	Group 2 (p_2)	Group 1 (O_1)	Group 2 (O_2)	RR	OR
1	0.001	0.003	0.002	0.003	3	3.01
2	0.01	0.03	0.01	0.03	3	3.06
3	0.02	0.06	0.02	0.06	3	3.13
4	0.05	0.15	0.05	0.18	3	3.35
5	0.10	0.30	0.11	0.43	3	3.86
6	0.15	0.45	0.18	0.82	3	4.64
7	0.20	0.60	0.25	1.50	3	6.00
8	0.25	0.75	0.33	3.00	3	9.00
9	0.30	0.90	0.43	9.00	3	21.0
10	0.33	0.99	0.49	99.0	3	2101.0

OR is similar to RR when the risk/probability is low

RR vs OR



JAMA. 1998;280(19):1690-1691. doi:10.1001/jama.280.19.1690

Case-control study: OR required

	Disease present	Disease absent	total
male	32	17*30	49
female	118	127*30	245
total	150	144*30	294

Suppose: sample has three times more cases (diseased) than the population

OR independent of prevalence of disease, i.e. how often disease occurs in population

Case-control study: OR

	Disease present	Disease absent	total
male	32	17	49
female	118	127	245
total	150	144	294

	Disease present	Disease absent	total
male	32	510	542
female	118	3810	3928
total	150	4320	4470

30 Times

Initial:

RR: (32/49) / (118/245) = 1.4
OR: (32/17) / (118/127) = 2.0

30 Times as many controls:

RR: (32/542) / (118/3928) = 1.97
OR: (32/510) /(118/3810) = 2.0

In case-control design: researcher determines ratio case/control
RR is affected by this choice, OR is NOT → OR in case-control design

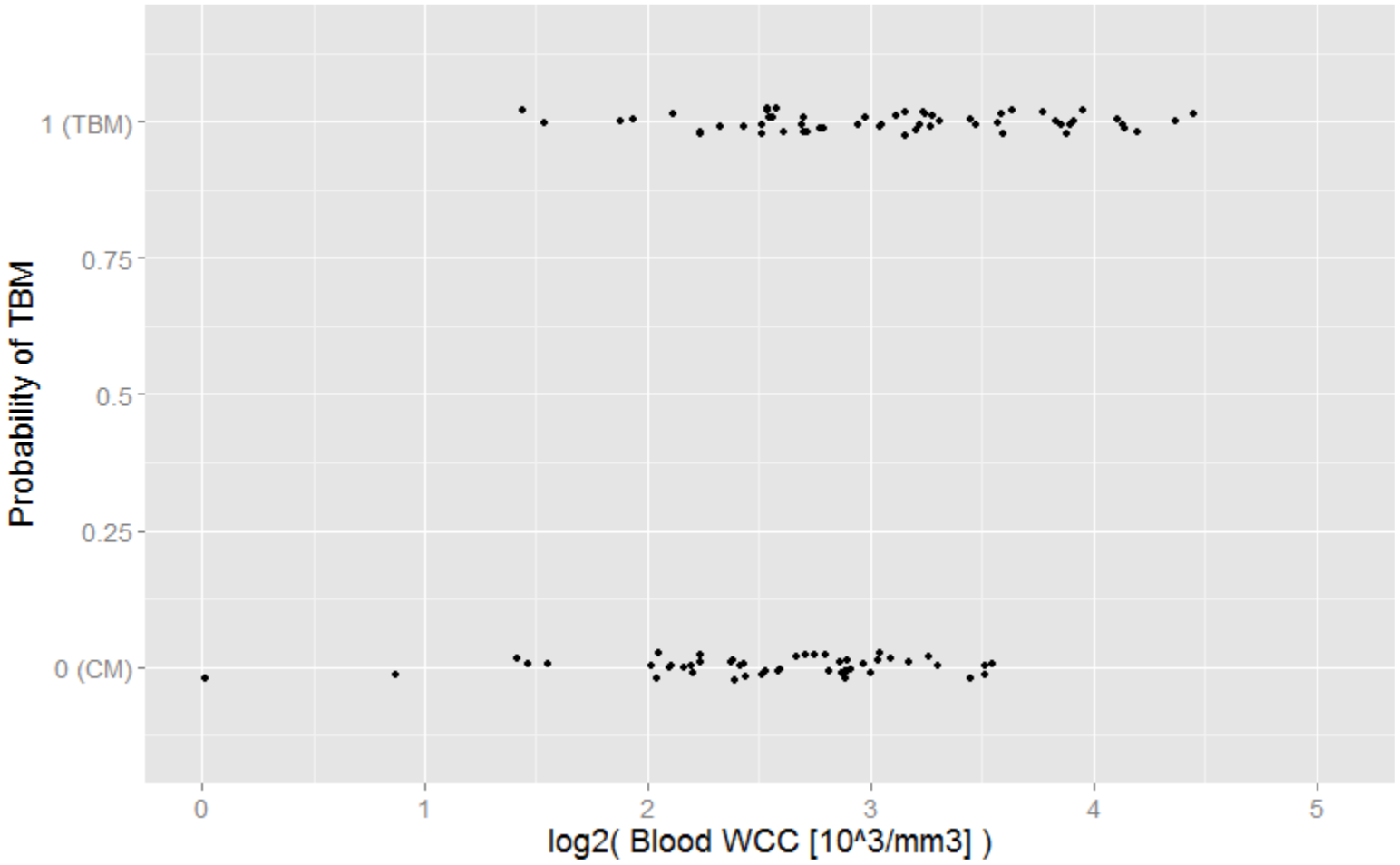
Logistic regression

Setting

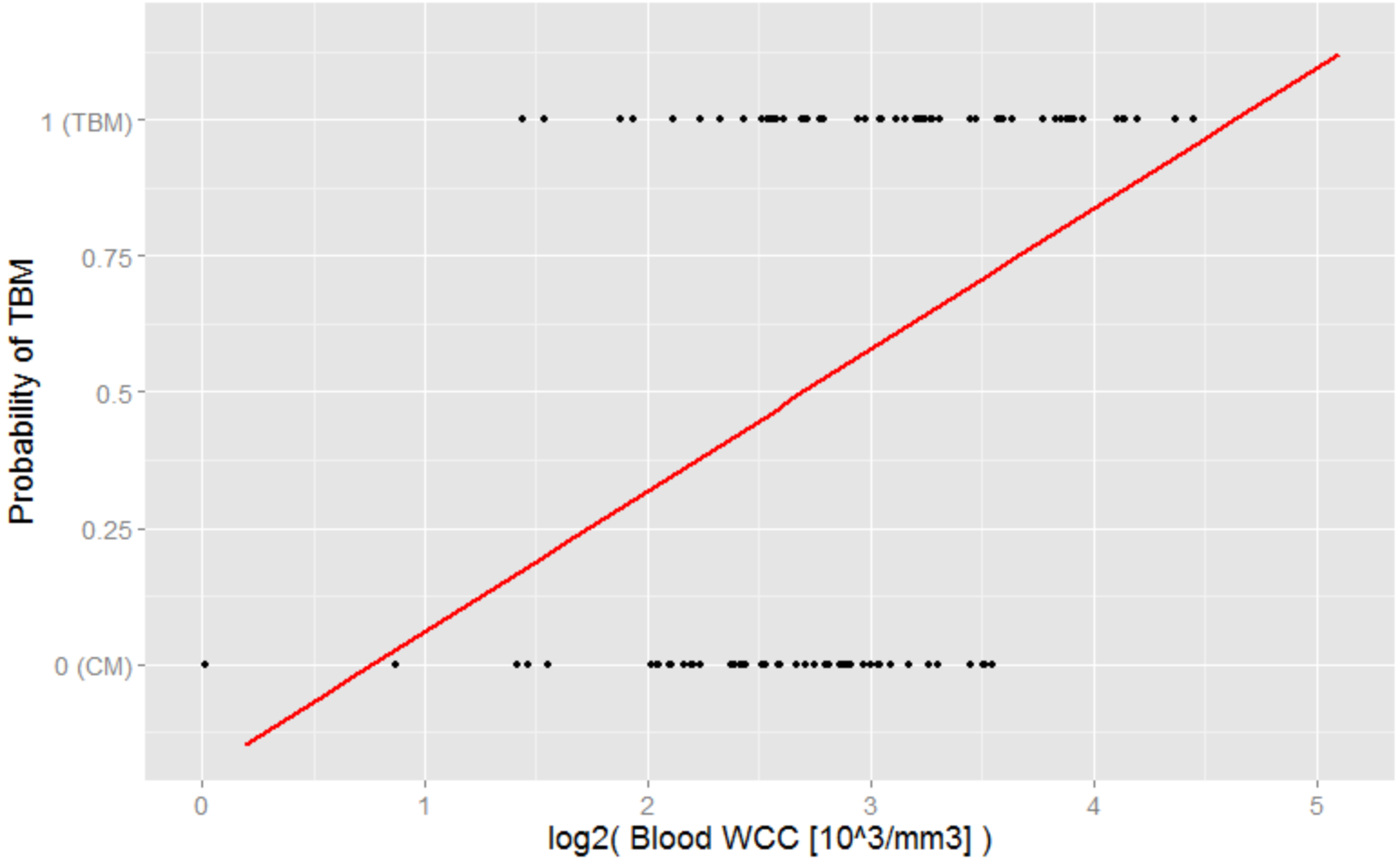
- Binary outcome variable Y taking on values 0 and 1
- A single covariable x
- Question
  - Is Y associated with x?
  - More precisely: How does the probability P(Y=1) depend on x?

- Example
  - 108 HIV+ subjects with non-missing WCC in blood
  - CM (cryptococcal meningitis) (n=58) and TBM (n=50)
  - Outcome Y: diagnosis (0="CM", 1="TBM")
  - Covariable x: WCC in blood (log2-transformed)

Raw data


1 (TBM)
0.75
0.5
0.25
0 (CM)
Probability of TBM
0
1
2
3
4
5
log2(Blood WCC [10^3/mm3])

Raw data with linear regression fit: A good idea?



The logistic regression model

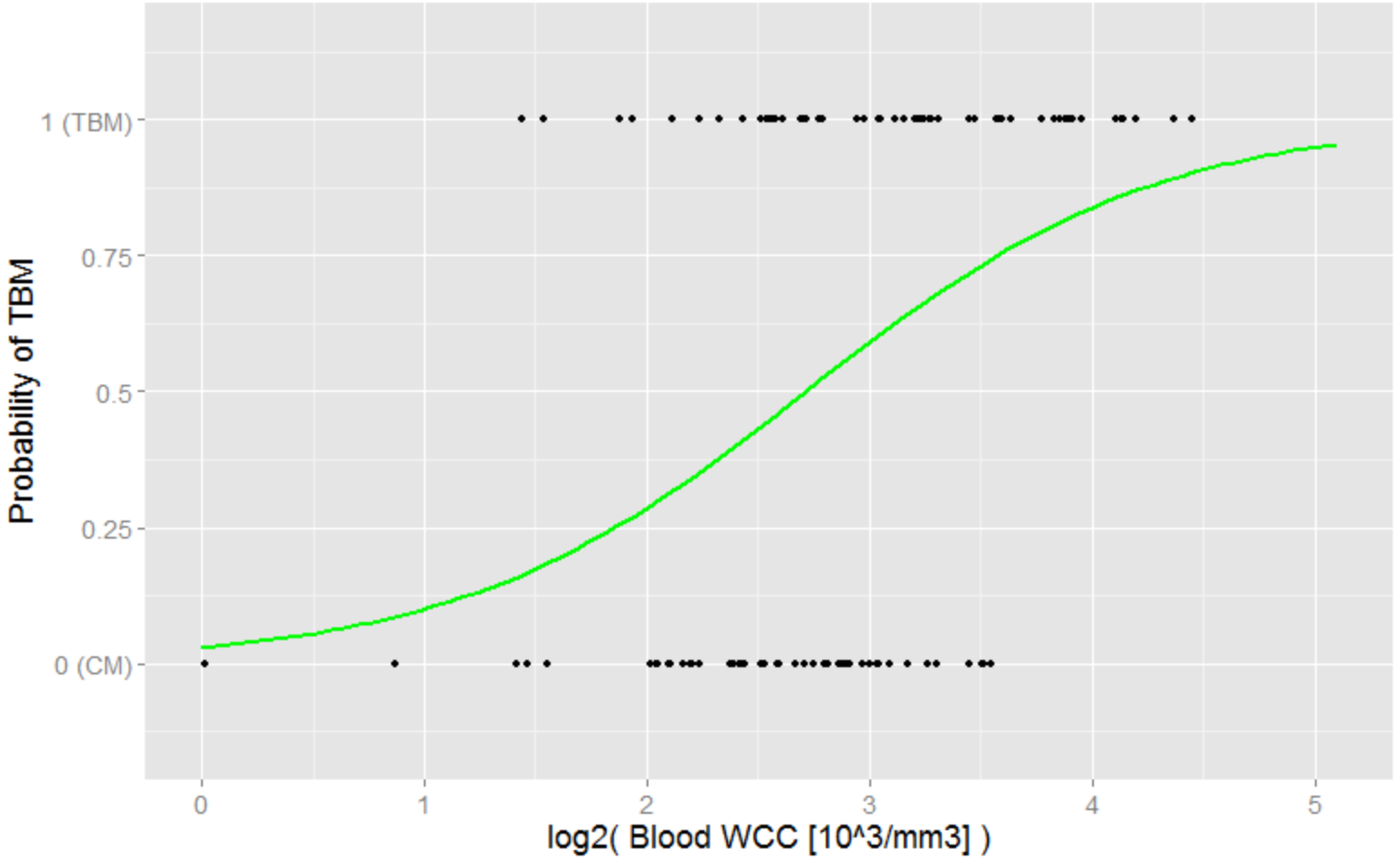
- Linear regression model
  - Y = a + bx + E
  - Not a good idea if Y only has values 0 and 1

- Logistic regression
  - A model for E(Y)=P(Y=1) (just like linear regression)
  - 0 ≤ P(Y=1) ≤ 1 → fit probability with an S-shaped curve

$$\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \mathrm{a} + \mathrm{bx}$$

$$P(Y=1) = \frac{\exp(a+bx)}{1+\exp(a+bx)}$$

Raw data with logistic regression fit



3 ways to write logistic regression model

- Model

$$P(Y=1)=\frac{\exp(a+bx)}{1+\exp(a+bx)}$$

- Model formulation via log odds (natural log, also written as ln)

$$\ln(\text{odds}(\text{Y}=1))=\ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right)=\text{a}+\text{bx}$$

logit(Y=1) = a + bx

- Model formulation using odds

$$\text{odds}(Y=1)=\frac{P(Y=1)}{1-P(Y=1)}=\exp(\text{a}+\text{bx})$$

Binary covariable (“two groups”)

- New treatment (or vaccine)
- Outcome: deceased within 12 months after start treatment
- Logistic model with 1 covariable (treatment)
- Coding treatment: 1 / 0

  X=1: new treatment; x=0: placebo

Regression model

Model formulation:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta * Treatment$$

New treatment

$$\ln\left(\frac{p_1}{1-p_1}\right) = \alpha + \beta$$

$$\ln\left(\frac{p_0}{1-p_0}\right) = \alpha$$

Placebo

$$\ln\left(\frac{p_1}{1-p_1}\right) - \ln\left(\frac{p_0}{1-p_0}\right) = \beta$$

$$\ln\left(\frac{p_1}{1-p_1} \Big/ \frac{p_0}{1-p_0}\right) = \beta = \ln \text{odds ratio}$$

$$\text{odds ratio} = e^{\beta}$$

"log-odds"
"logit"

Interpretation coefficients

- Model = $\ln(\frac{p}{1-p}) = \alpha + \beta * treatment$
- $e^{\beta_{\textbf{treatment}}}$ = $\frac{\text{odds } \dagger \text{ treatment}}{\text{odds } \dagger \text{ placebo}}$
- Interpretation of intercept α:
  - log-odds of Y=1 if x=0 (placebo)
    → exp(α) is the odds of Y=1 if x=0
  - Often omitted from logistic regression summaries

Formulation in R

- Influenza; covariable vaccine (1=yes, 0=no)

```
fit <- glm(influenza ~ vaccine, data=influenza.trial, family=binomial)
library(gtsummary)
tbl_regression(fit, exponentiate=TRUE)
```

Characteristic	OR	95% CI	p-value
vaccine	0.16	0.09, 0.27	<0.001

OR of getting influenza for vaccine vs. placebo

- Note: can also be analysed as 2x2 table

Continuous covariable

- Model

$$\ln(\text{odds}(Y=1)) = a + bx$$

- Interpretation of intercept a:
  - log-odds of Y=1 if x=0 → exp(a) is the odds of Y=1 if x=0

- Interpretation of slope b:
  - log-odds comparing Y=1 for x+1 vs. x

    (ln(odds(Y=1|covariable=x+1)) - (ln(odds(Y=1|covariable=x)) = b
  - exp(b) represents the change in the odds of Y=1 by increasing x by 1 unit

Fitting logistic regression model in R

```
fit <- glm(tbm ~ log2.bldwcc, data=cm.tbm.hiv, family=binomial)
tbl_regression(fit, exponentiate=TRUE, intercept=TRUE)
```

Characteristic	OR[1]	95% CI[1]	p-value
(Intercept)	0.03	0.00, 0.19	<0.001
log2.bldwcc	3.60	1.92, 7.47	<0.001

exp(a) and exp(b)

95% CI for exp(a) and exp(b)

p-value of null hypothesis that true slope b=0 (or, equivalently, that OR=exp(b)=1).

Reporting of results

- Example Table:
  Logistic regression for the probability of having TBM (compared to CM) depending on blood WCC [$10^3/mm^3$, log2-transformed].

Variable	OR	95% CI	p-value
Blood WCC [for each two-fold increase]	3.60	1.92-7.47	<0.0001

(Note that an increase by +1 in log2(bldwcc) corresponds to a 2-fold increase in bldwcc.)

Prediction in R

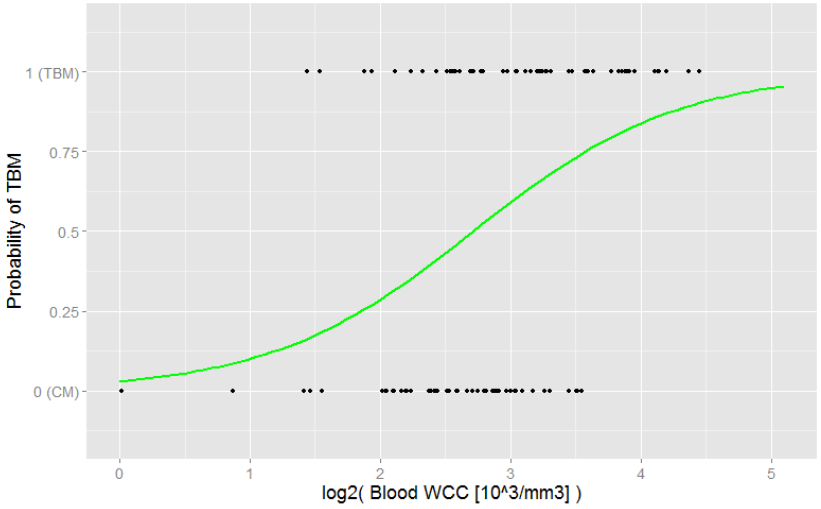
- What are the predicted probabilities of TBM for two subject with a blood WCC of 4 or 16 [x10^3/mm3], i.e. log2.bldwcc=2 or 4, respectively, based on the univariable logistic regression model?

```
fit <- glm(tbm ~ log2.bldwcc, data=cm.tbm.hiv, family=binomial)
newdata <- data.frame(log2.bldwcc=c(2,4))
predict(fit, newdata, type="response")
          1           2
0.2860918   0.8385559
```

- We can also do this for any value of WCC and obtain a plot as in slide 17

Logistic regression

y~x




```
fit <- glm(tbm~log2.bldwcc,
           data=cm.tbm.hiv,
           family=binomial)
```

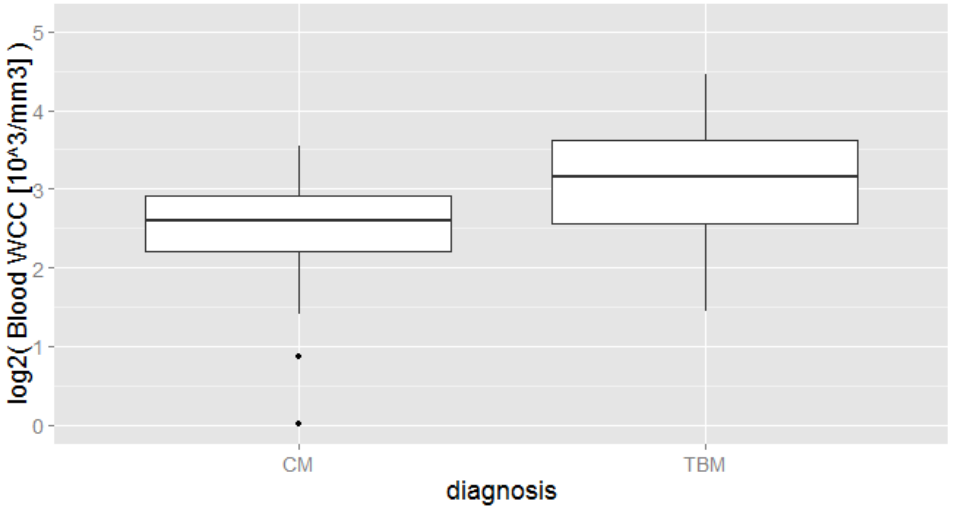
Result:

OR (95% CI) of TBM for each 2-fold increase in bldwcc: 3.60 (1.92 to 7.47).

Test for no effect (OR=1): p<0.0001

t-test (reverses causal order)

x~y (assuming x continuous)




```
t.test(log2.bldwcc~tbm,
       data=cm.tbm.hiv)
```

Result:

Estimate (95% CI) for difference in log2.bldwcc: 0.59 (0.32 to 0.85)

Test for no difference: p<0.0001

Multivariable logistic regression

Univariable versus multivariable logistic regression

- Univariable – a single covariable
- Multivariable – more than one covariable
  - Explore: to assess “independent” covariable effects (after “controlling” for other covariables)
  - Predict: to combine multiple covariables in a prognostic model
  - Explain: to adjust a covariable/exposure of main interest for potential confounders
- Extension from univariable to multivariable for logistic regression is analogous to linear regression.
- Same for interactions, non-linear trends

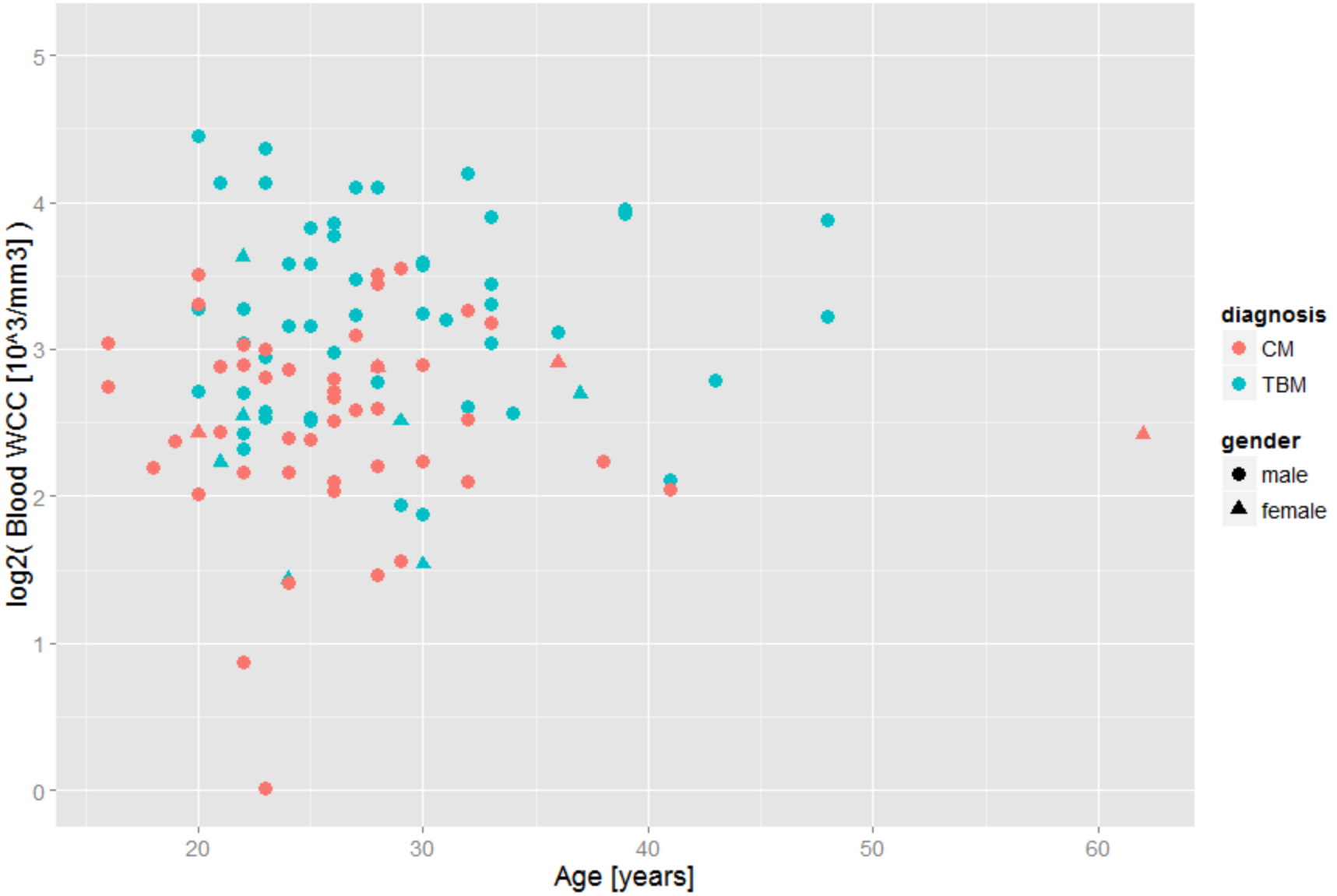
Multivariable logistic regression model

- Model for 3 covariables (formulation using log odds)

$$\log(\text{odds}(Y=1)) = \log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = a + b_1 x_1 + b_2 x_2 + b_3 x_3$$

- Example: CM/TBM data
  - Y: diagnosis (1="TBM", 0="CM")
  - Covariable:
    - x1: WCC in blood (log2-transformed)
    - x2: age (years)
    - x3: gender

Example – Graphical display



Multivariable logistic regression in R

```
fit <- glm(tbm ~ log2.bldwcc + I((age-20)/10)+ gender,
           data = cm.tbm.hiv, family = binomial)
```

	log.OR	OR	lower.CI	upper.CI	p.value
(Intercept)	-4.08	0.02	0.00	0.12	NA
log2.bldwcc	1.39	4.00	2.06	8.60	0.0000
I((age - 20)/10)	0.31	1.37	0.75	2.68	0.3214
genderfemale	0.98	2.66	0.65	12.60	0.1754

OR of having TBM for each +1 increase in log2.bldwcc (after adjusting for all other covariables).

p-value of the null hypothesis that the true regression coef (log.OR) b_1 is 0, i.e. that the true OR is 1, i.e. no association with outcome (after controlling for other covariables)

Presentation of results; comparison

- Univariable logistic regression models (separate models for each covariable):

Variable	OR	95% CI	p-value
Blood WCC [x 2]	3.60	1.92-7.47	<0.0001
Age [+10 years]	1.49	0.85-2.77	0.16
Female gender	1.26	0.38-4.53	0.71

- Multivariable logistic regression model

Variable	OR	95% CI	p-value
Blood WCC [x 2]	4.00	2.06-8.60	<0.0001
Age [+10 years]	1.37	0.75-2.68	0.32
Female gender	2.66	0.67-12.60	0.18

Example – 3 step recipe to test for a potential quadratic effect of log2.bldwcc

1. Fit original model without quadratic effect (as before):

```
fit <- glm(tbm ~ log2.bldwcc + age + gender,
           data=cm.tbm.hiv, family=binomial)
```

2. Fit a more complex model with quadratic effect:

```
fit.quad <- glm(tbm ~ poly(log2.bldwcc,degree=2) + age +
           gender, data=cm.tbm.hiv, family=binomial)
```

3. Do a statistical test to compare whether model 2. is "better" than model 1.

```
anova(fit,fit.quad,test="Chisq")
Model 1: tbm ~ log2.bldwcc + age + gender
Model 2: tbm ~ poly(log2.bldwcc, degree = 2) + age + gender
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       103     125.95
2       102     123.85  1   2.0952   0.1478
```

← No convincing evidence that a quadratic term improves model.

Example – 3 step recipe to test for a potential interaction between log2.bldwcc and gender

1. Fit original model without interaction (as before):

```
fit <- glm(tbm ~ log2.bldwcc + age + gender,
                 data=cm.tbm.hiv, family=binomial)
```

2. Fit a more complex model with an interaction term:

```
fit.ia <- glm(tbm ~ log2.bldwcc + age + gender +
        log2.bldwcc:gender, data=cm.tbm.hiv, family=binomial)
```

3. Do a statistical test to compare whether model 2. is "better" than model 1.

```
anova(fit, fit.ia, test = "Chisq")
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1 103 125.95 2 102 121.13 1 4.8193 0.02814 *
```

Fairly strong evidence that log2.bldwcc and gender interact (i.e. that WCC affects the probability of TBM differently in males and females).
A detailed analysis shows that WCC were higher for TBM than CM in males but not in females. However, this should be interpreted with caution, as the dataset has only 12 females.

Limitations on the number of covariables

- If the number of covariables *p* is large, there is a risk that the regression fits adapts too much to the examined dataset (but not the population)
- This phenomenon is called over-fitting and implies that results will not replicate in future validation studies
- Rule of thumb: To avoid over-fitting, the number of covariables considered should be less than the *critical sample size* m divided by 10: $p \leq m/10$
  - The *critical sample* size is min(# with outcome 0,# with outcome 1)
  - If sample size is n= 200 (120 with Y=0, 80 with Y=1 → m=80) → one should not consider >8 covariables

→ Carefully select your covariables!

Modeling strategies for logistic regression

- Define covariables (and outcomes) a-priori
- Explore both univariable and multivariable logistic regression
- Assess model assumptions
  - Consider log-transformation for skewed covariables
  - Explore potential non-linear effects
  - If there are clinically plausible interactions, test for them
- Try to interpret the fitted model – does it make sense?
- Supplement the logistic regression by descriptive and graphical analyses.
- Get the help of a statistician for the development of complex logistic models.

Logistic and linear regression

- Logistic regression is an extension of linear regression to the setting of a binary outcome Y (coded as 0/1)
- Logistic regression is probably the most frequently applied regression method for medical data.
- Many similarities to linear regression exist but logistic regression is more complicated:
  - In linear regression, the outcome Y is directly modeled “linearly”
  - In logistic regression, P(Y=1) is modeled “non-linearly”
  - Regression diagnostics is more difficult for logistic regression (not covered)