

## Lecture 16

1. Risk and odds
2. Regression with binary response
3. Proc Logistic

*Logistic regression using the SAS system: theory and application* by Paul Allison, 2001, Wiley-SAS

*Categorical Data Analysis Using the SAS System, 2nd ed.* by Stokes, Davis, and Koch, 2009, SAS Institute

“Proc Logistic: Traps for the Unwary,” by P.L. Flom (on course website)

1

### **2×2 Tables: Relative Risk, Odds Ratio**

Want to compare rate of event between two groups:

	<i>Event</i>	<i>No Event</i>
<i>Group 1</i>	7	93
<i>Group 2</i>	1	99

Usually interested in row percent = **rate of events**

2

Risk = rate = proportion of subjects with event

= row percent (when event is column)

	<b>Event</b>	<b>No Event</b>	risk
<b>Group 1</b>	A	B	$A/(A + B)$
<b>Group 2</b>	C	D	$C/(C + D)$

	<b>Event</b>	<b>No Event</b>	risk
<b>Adults</b>	7	93	7%
<b>Children</b>	1	99	1%

Sample proportion in adults is  $\hat{p}_a = 7/100 = 7\%$

in children, sample proportion is  $\hat{p}_c = 1/100 = 1\%$ .

3

### Two different null hypotheses for proportions

Two ways to compare population proportions  $p_a$  and  $p_c$ :

1. Test whether the **risk difference** is zero,  $H_0: p_a - p_c = 0$

*Z*-test, based on  $(\hat{p}_a - \hat{p}_c)$ , is equivalent to chi-square test

Alternative test of risk differences when some cells have small counts:

*Fisher's exact test*.

2. Test whether the **risk ratio**, or **relative risk** is one:

$$H_0: \frac{p_a}{p_c} = 1$$

Null value for differences is 0. Null value for ratios is 1.

4

Proc FREQ;

tables age \* event\_status / nopercnt nocol **chisq** **relrisk** ;

age	event_status		
Frequency			
Row Pct	Event	No Event	Total
Adult	7	93	100
	7.00	93.00	
Peds	1	99	100
	1.00	99.00	
Total	8	192	200

5

Statistic	DF	Value	Prob
<b>Chi-Square</b>	1	4.6875	<b>0.0304</b>

Estimates of the **Relative Risk** (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	7.4516	0.8995	61.7288
Cohort (Col1 Risk)	7.0000	<b>0.8772</b>	<b>55.8565</b>
Cohort (Col2 Risk)	0.9394	<b>0.8871</b>	<b>0.9948</b>

Risk difference significant? Either relative risk significant?

6

WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test

```

-----
Cell (1,1) Frequency (F)          7
Left-sided Pr <= F                0.9966
Right-sided Pr >= F              0.0324

Table Probability (P)             0.0291
Two-sided Pr <= P                0.0649
  
```

Is risk difference really significant?

7

Another summary of event data is the **odds**.

$$\text{odds} = \frac{\text{number of events}}{\text{number without event in sample}}$$

	<i>Event</i>	<i>No Event</i>	odds	risk
<b>Group 1</b>	<i>A</i>	<i>B</i>	<i>A/B</i>	<i>A/(A + B)</i>
<b>Group 2</b>	<i>C</i>	<i>D</i>	<i>C/D</i>	<i>C/(C + D)</i>

Relating odds to risk:

$$\text{odds} = \frac{\text{number of events}}{\text{number without event}} = \frac{\text{number of events} / n}{\text{number without event} / n} = \frac{\hat{p}}{1 - \hat{p}}$$

For rare events,  $\hat{p} \approx 0$  and so the denominator is almost 1, and  $\text{odds} \approx \text{risk}$ .

	<i>Event</i>	<i>No Event</i>	odds	risk
<b>Group 1</b>	A	B	A/B	A/(A + B)
<b>Group 2</b>	C	D	C/D	C/(C + D)

Odds are compared only by ratio, never by difference.

**Odds ratio** is the odds in the top row divided by odds in the bottom row, which simplifies to  $AD/BC$ .

Test whether population odds ratio is one,  $H_0: OR = 1$

by checking whether the 95% confidence interval covers 1.

9

```
Proc FREQ;
  tables age * event_status / nopercnt nocol relrisk ;
```

age	event_status		
Frequency			
Row Pct	Event	No Event	Total
Adult	7	93	100
	7.00	93.00	
Peds	1	99	100
	1.00	99.00	
Total	8	192	200

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	7.4516	0.8995	61.7288

## Binary responses

Event/no-event, or 0/1 responses are **binary responses**.

Set-up: one trial in which  $Y$  takes values 1=event/0=no event.

$$P[\text{event}] = P[Y = 1] = \pi = \text{population event rate}$$

$$P[\text{no event}] = P[Y = 0] = 1 - \pi$$

$Y$  has **Bernoulli** distribution (after Jakob Bernoulli, 1654–1705).

$$\text{mean of } Y = \pi, \quad \text{standard deviation of } Y = \sqrt{\pi(1 - \pi)}.$$

*SD is a function of the mean, unlike Normal distribution.*

11

**Grouped binary responses:** If many trials under done with same chance of event, such as flipping a coin  $n$  times, then the results are often grouped as  $X = k$  events in  $n$  trials, or  $X = k/n$ .

Each of the  $n$  trials is a Bernoulli random variable,  $Y_i$ , and we can write  $X$  as:

$$X = Y_1 + \cdots + Y_n = k.$$

The random variable  $X$  is the *sum* of  $n$  independent Bernoulli random variables  $\{Y_i\}$ , and has a **binomial distribution**  $\mathcal{B}(X, \pi, n)$ .

12

## Binomial distribution $\mathcal{B}(X, \pi = 0.5, n = 6)$

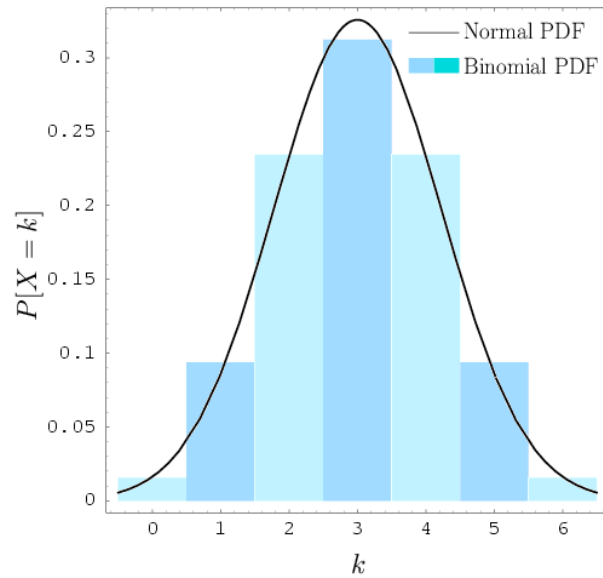


illustration from Wikipedia

The Normal curve is drawn here because, for large  $n$ , it can be used to approximate the binomial.

13

## Regression with binary responses

*Continuous* response  $y$ , regression models *mean* of  $y$  as a function of predictors  $x$

$$\mu_Y(x) = \beta_0 + \beta_1 x$$

Binary (0/1) response  $y$ : regression models *mean* of  $y$  as a function of predictors  $x$

$$\pi(x) = f(\beta_0 + \beta_1 x)$$

Many choices for  $f$ : logistic, probit, log-binomial, Poisson

but not the identity function

14

## Example: Obesity in NHANES 2004

NHANES 2004 data for children and adults people under age 50 ( $n = 6116$ )

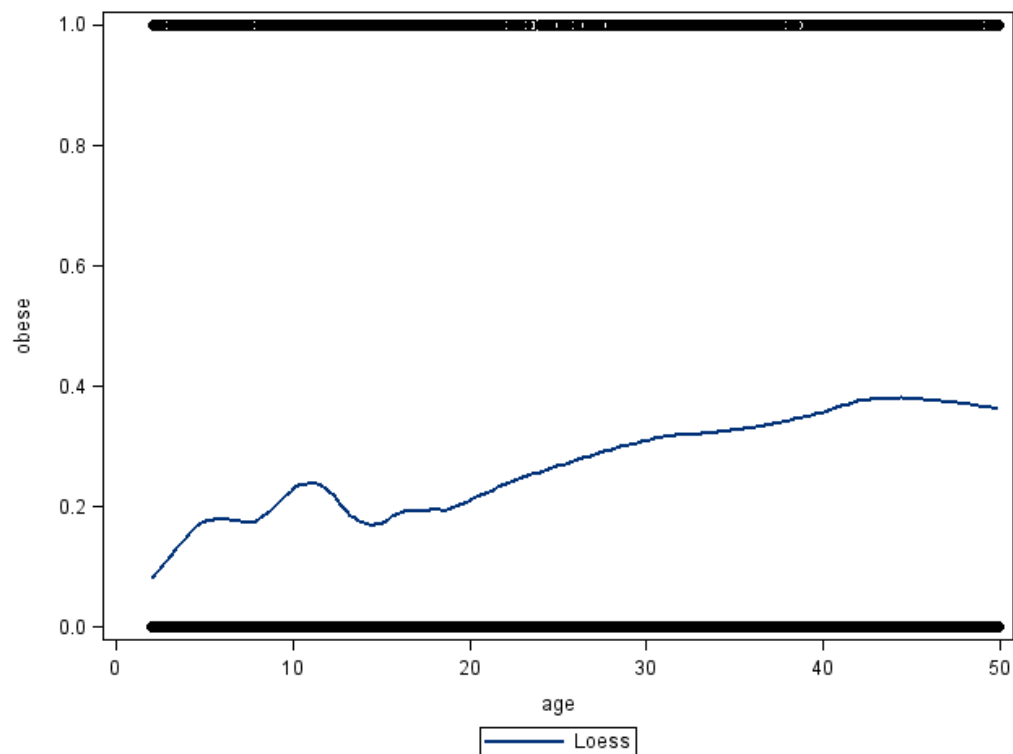
Event = obesity, defined as  $BMI \geq 30$ , or  $\geq 95$ th percentile for children

Consider rate of obesity as function of age:  $P[\text{obese} | \text{age}] = \pi(\text{age})$

Graph data, use smoother to estimate  $\pi(\text{age})$  without assuming shape

```
ODS graphics on;  
Proc SGplot data=under50;  
  loess y = obese x = age ;  
run;  
ODS graphics off;
```

15



16

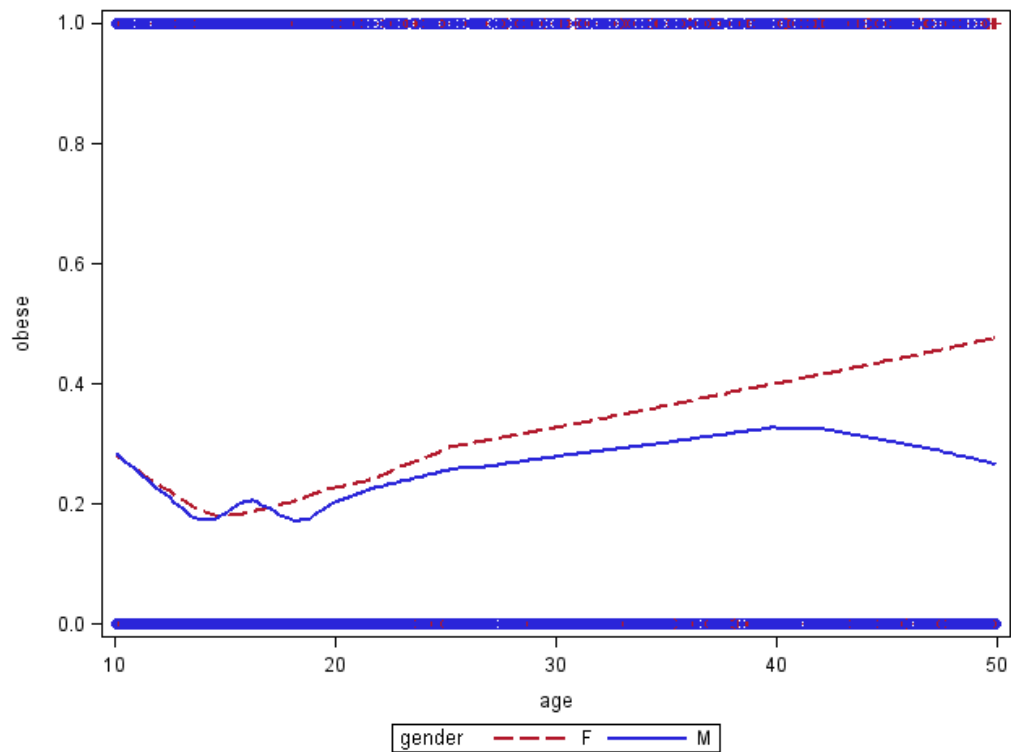
Smooth separately for each gender

```
ODS graphics on;  
Proc SGplot data=under50;  
  where age > 10.0 ;  
  loess y = obese x = age / group = gender ;  
run;  
ODS graphics off;
```

LOESS = local linear regression, doesn't like to estimate zero

Sometimes need to restrict range.

17



18

## Logistic link between mean and predictors

$$\text{mean} = P[\text{obese} | \text{age}] = \pi(\text{age}) = \frac{\exp(\beta_0 + \beta_1 \text{age})}{1 + \exp(\beta_0 + \beta_1 \text{age})} = \frac{1}{1 + \exp(-\beta_0 - \beta_1 \text{age})}$$

Logistic curve on the probability scale.

Equivalent:

$$\log\left(\frac{\pi(\text{age})}{1 - \pi(\text{age})}\right) = \beta_0 + \beta_1 \text{age}$$

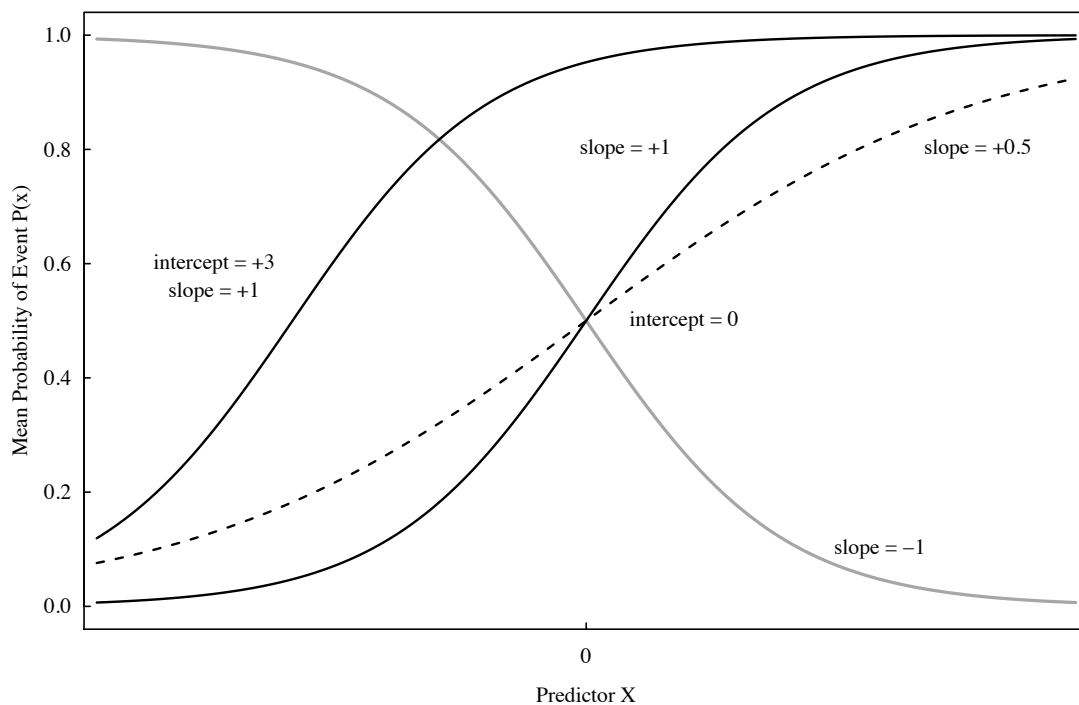
Function on left is **log odds** or **logit** because

$$\frac{\hat{\pi}}{1 - \hat{\pi}} = \frac{\text{number with event} / n}{\text{number without event} / n} = \frac{\text{number with event}}{\text{number without event}} = \text{odds}$$

Linear on the log odds (logit) scale.

19

Logistic curves on probability scale  $\pi(x) = \text{intercept} + (\text{slope})x = \beta_0 + \beta_1 x$

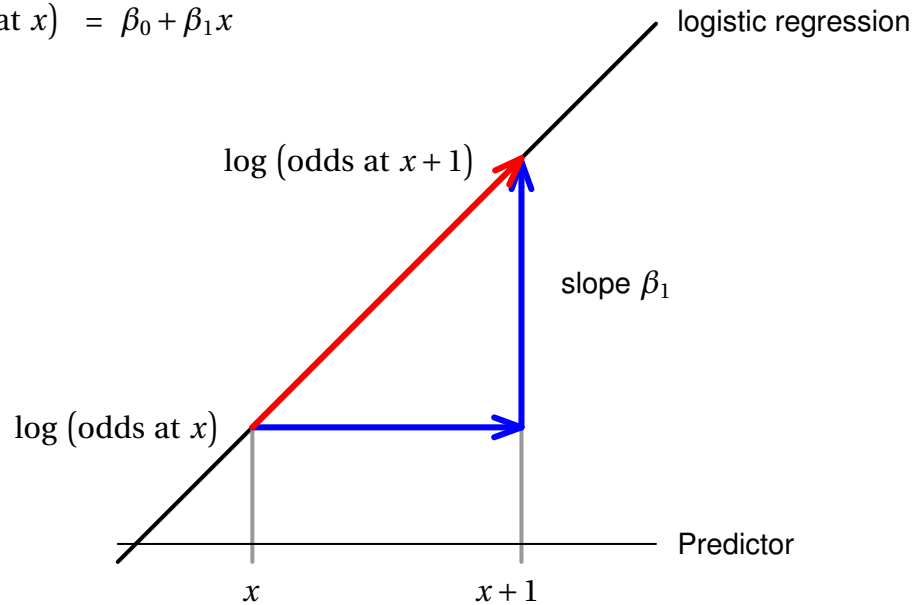


20

## Interpreting slope in logistic regression

On log-odds scale, logistic function is a line:

$$\log(\text{odds at } x) = \beta_0 + \beta_1 x$$



21

Slope is change in  $\log(\text{odds})$  for unit change in  $x$

$$\text{slope } \beta_1 = \log(\text{odds at } x+1) - \log(\text{odds at } x)$$

On log scale,  $\log A - \log B = \log(A/B)$ , so

$$\text{slope } \beta_1 = \log(\text{odds at } x+1) - \log(\text{odds at } x) = \log\left(\frac{\text{odds at } x+1}{\text{odds at } x}\right)$$

Apply exponential function as inverse of log:

$$\exp(\text{slope } \beta_1) = \frac{\text{odds at } x+1}{\text{odds at } x}$$

exponential of slope = odds ratio for unit increase in  $x$

22

## Proc Logistic

```
Proc Logistic data=under50;  
  model obese = age; model format like Proc GLM
```

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.8317	0.0594	949.9665	<.0001
age	1	-0.0302	0.00227	176.0425	<.0001

### Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
age	0.970	0.966 0.975

$\exp(-0.0302) = .9702515$  Correct, but does this make sense?

23

First part of the output:

### The LOGISTIC Procedure

```
Data Set                WORK.ADULT  
Response Variable       obese  
Number of Response Levels 2  
Model                   binary logit  
Optimization Technique  Fisher's scoring
```

```
Number of Observations Read    6116  
Number of Observations Used    6116
```

### Response Profile

Ordered Value	obese	Total Frequency
1	0	4694
2	1	1422

Probability modeled is obese=0.

24

There is also a warning in the log

```
NOTE: PROC LOGISTIC is modeling the probability that obese=0. One way to change  
      this to model the probability that obese=1 is to specify the response  
      variable option EVENT='1'.
```

```
NOTE: Convergence criterion (GCONV=1E-8) satisfied.
```

```
NOTE: There were 6116 observations read from the data set WORK.UNDER50.
```

Proc Logistic models probability of *no event* by default.

Opposite of what everyone expects.

25

Two fixes:

1. Use the `descending` option to make SAS fit the probability that  $y = 1$ . Works in both Logistic and Proc Genmod (which also fits logistic regression).

```
Proc Logistic descending data=under50;  
      model obese = age;
```

2. Define the event in the `MODEL` statement. *Only works in Logistic.*

```
Proc Logistic data=under50;  
      model obese (event = '1') = age;
```

Clearer code, but unfortunately doesn't work in Proc Genmod.

26

```
Proc Logistic data=under50;
  model obese (event = '1') = age;
```

NOTE: PROC LOGISTIC is modeling the probability that obese=1.

NOTE: Convergence criterion (GCONV=1E-8) satisfied.

NOTE: There were 6116 observations read from the data set WORK.UNDER50

27

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.8317	0.0594	949.9665	<.0001
age	1	0.0302	0.00227	176.0425	<.0001

#### Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
age	1.031	1.026 1.035

Odds of obesity increase by a factor of 1.03 per year for people under 50 (95% confidence interval 1.026–1.035).

28

## Changing units for odds ratios

To get odds ratios for a 10-year change in age:

```
Proc Logistic descending data=under50;
  model obese = age / rsquare CLodds = PL;
  units age = 10.0 ;
```

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.8317	0.0594	949.9665	<.0001
age	1	0.0302	0.00227	176.0425	<.0001

Effect	Odds Ratio Estimates		
	Point Estimate	95% Wald Confidence Limits	
age	1.031	1.026 1.035	

Profile Likelihood Confidence Interval for Odds Ratios				
Effect	Unit	Estimate	95% Confidence Limits	
age	10.0000	1.352	1.293 1.414	

29

Two useful options to the model statement in Proc Logistic:

```
Proc Logistic descending;
  model obese = age / CLodds = PL Rsquare ;
```

Default (Wald) confidence intervals for OR are  $\exp(\hat{\beta} \pm 1.965 * SE(\hat{\beta}))$

`CLodds = PL` “profile likelihood confidence intervals” are more accurate than Wald for small samples. Gives same answer for large samples.

`Rsquare` gives a version of R-square from linear regression

Maximum possible value of generalized  $R^2$  is not 1.0 as for linear regression.

Max-rescaled R-Square divides by this maximum value to fix this.

30

R-Square 0.0282 Max-rescaled R-Square 0.0427

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age	1.031	1.026	1.035

Profile Likelihood Confidence Interval for Odds Ratios

Effect	Unit	Estimate	95% Confidence Limits	
age	1.0000	1.031	1.026	1.035

Profile Likelihood CI is identical here because sample size is large ( $n = 6116$ )

31

**CLASS variables in Proc Logistic**

```
proc logistic descending data=under50;
```

```
  class gender;
```

```
  model obese = age gender;
```

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.8326	0.0594	950.7355	<.0001
age	1	0.0302	0.00227	176.6801	<.0001
gender F	1	0.0779	0.0308	6.4109	0.0113

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age	1.031	1.026	1.035
gender F vs M	1.169	1.036	1.318

For age,  $\exp(0.0302) = 1.031$ , but for gender,  $\exp(0.0779) = 1.081$

32

The LOGISTIC Procedure

Model Information

Data Set WORK.ADULT  
Response Variable obese  
Number of Response Levels 2  
Model binary logit  
Optimization Technique Fisher's scoring

Class Level Information

Class	Value	Design Variables
gender	F	1
	M	-1

Default coding for CLASS variables is not the same as Proc GLM (0/1)

33

If you want to work with regression coefficients, then request 0/1 indicator variables.

```
proc logistic descending data=under50;  
  class gender / param = GLM ;  
  model obese = age gender
```

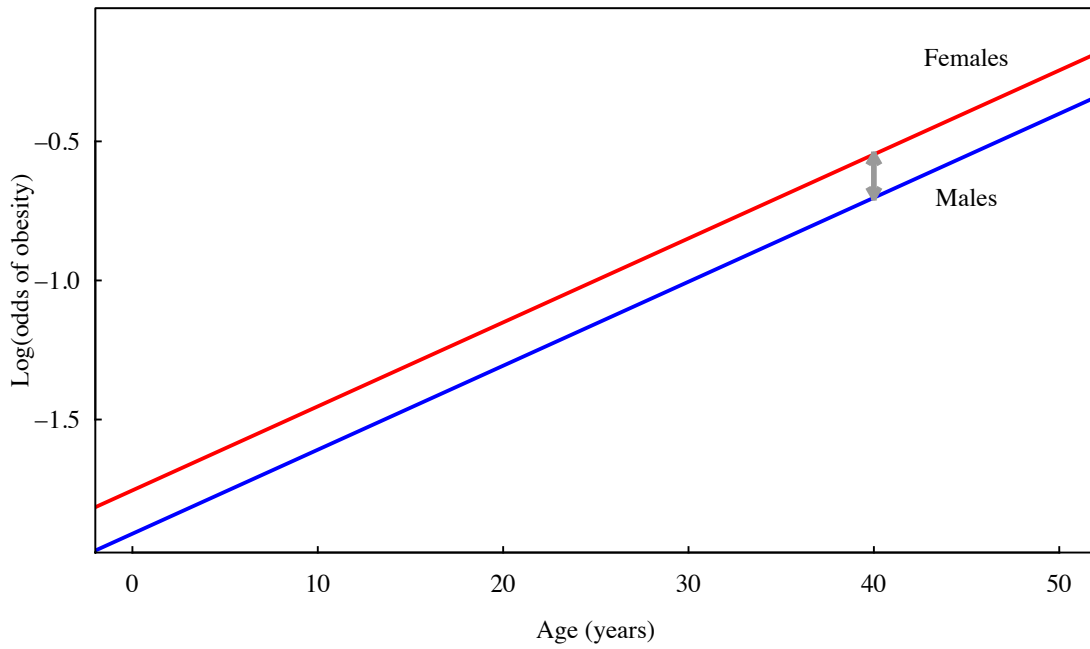
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.9105	0.0675	800.7992	<.0001
age	1	0.0302	0.00227	176.6801	<.0001
gender F	1	0.1558	0.0615	6.4109	0.0113
gender M	0	0	.	.	.

Effect	Odds Ratio Estimates		
	Point Estimate	95% Wald Confidence Limits	
age	1.031	1.026	1.035
gender F vs M	1.169	1.036	1.318

$\exp(0.1558) = 1.168592$

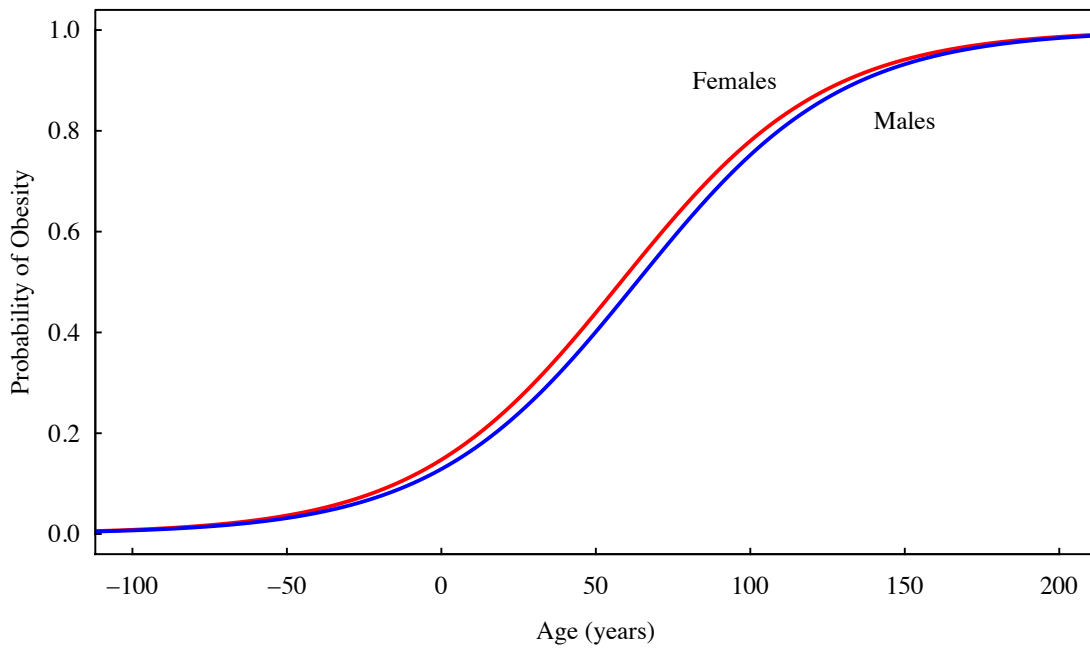
34

Logistic regression fits *linear* model on log(odds) scale. obese = age gender is parallel lines model. Distance between lines = log(odds ratio).



35

Parallel lines model on probability scale.



36

Check for interaction (*separate lines model*):

```
Proc Logistic descending data=under50;  
  class gender;  
  model obese = age gender age*gender ;
```

Surprise: get regression coefficients but no odds ratios.

37

Type 3 Analysis of Effects

Effect	DF	Chi-Square	Wald Pr > ChiSq
age	1	171.9228	<.0001
gender	1	2.0908	0.1482
age*gender	1	10.3290	0.0013

Analysis of Maximum Likelihood Estimates *reg coef*

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.8266	0.0595	942.9555	<.0001
age	1	0.0299	0.00228	171.9228	<.0001
gender female	1	-0.0860	0.0595	2.0908	0.1482
age*gender female	1	0.00733	0.00228	10.3290	0.0013

Proc Logistic does not calculate odds ratios for effects included in interaction terms:

**no odds ratios for age or gender**

38

## Why no odds ratio for terms in interaction?

```
Proc Logistic descending data=young;  
  class gender;  
  model obese = age gender age*gender / rsquare CLodds = PL;
```

### Analysis of Maximum Likelihood Estimates *reg coef*

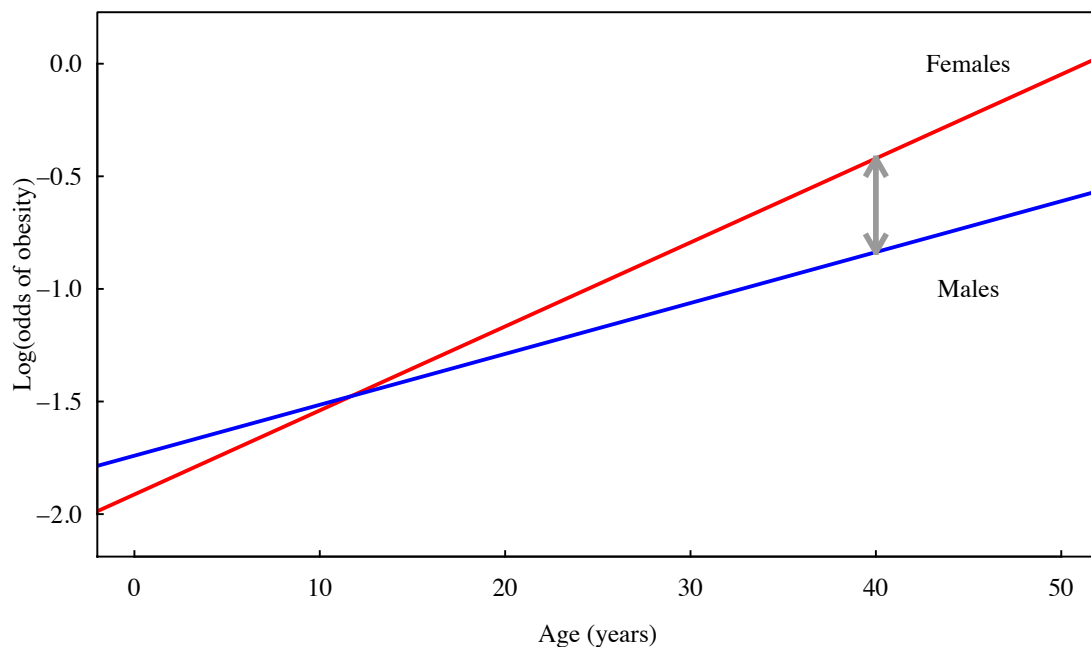
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.8266	0.0595	942.9555	<.0001
age	1	0.0299	0.00228	171.9228	<.0001
gender female	1	-0.0860	0.0595	2.0908	0.1482
age*gender female	1	0.00733	0.00228	10.3290	0.0013

Model fits separate line for each gender: 2 intercepts, 2 slopes

39

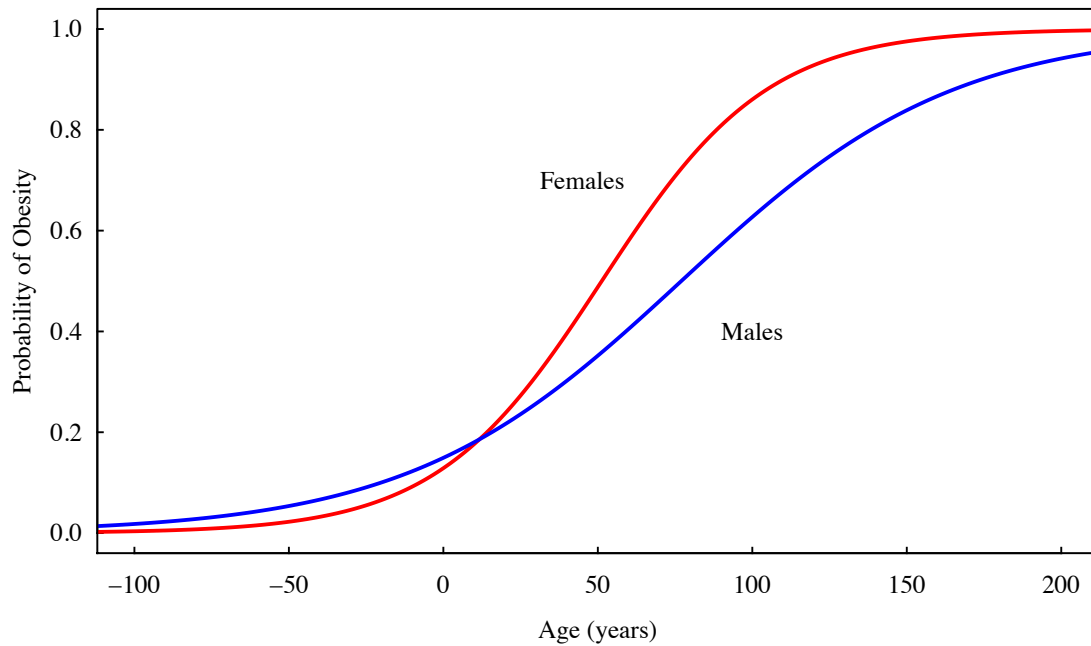
Separate line for each gender on log(odds) scale.

Distance between gender lines is different at every age: **no single gender effect**



40

Separate line for each gender on probability scale.



41

### ODDSRATIO statement

This is like LSmeans in Proc GLM: gives odds ratios for combinations of factor-levels in the interaction.

```
proc logistic descending data=under50;
  class gender;
  model obese = age gender age*gender / rsquare CLodds = PL;
  oddsratio age ;
  oddsratio gender / at (age=15 25 35 45);
```

42

## Wald Confidence Interval for Odds Ratios

Label	Estimate	95% Confidence Limits	
age at gender=F	1.038	1.032	1.044
age at gender=M	1.023	1.016	1.029

*separate age odds ratios for each gender)*

gender F vs M at age=15	1.049	0.915	1.204
gender F vs M at age=25	1.215	1.074	1.374
gender F vs M at age=35	1.407	1.192	1.661
gender F vs M at age=45	1.629	1.286	2.063

*comparisons of genders adjusted for specific ages*

43

### Summary: surprises in Proc Logistic

1. Fits probability of *no event* by default. Opposite of what everyone expects.
2. Codes class variables using +1/ - 1 rather than 0/1. Makes regression coefficients difficult to interpret.
3. Odds ratios are main effects, so no odds ratios for terms involved in interactions. Use `oddsratio` statement.
4. More: see “Proc Logistic: Traps for the Unwary,” by P.L. Flom (on course website)

44