PhD course in Basic Biostatistics – Day 7

Erik Parner, Department of Biostatistics, Aarhus University©

Binary regression models

Odds and odds ratios (repetition from Day 4)

Logistic regression and odds ratio

A simple logistic regression model
 Post term delivery and age of the woman

Comparing two groups after adjustment for a covariate
 Post term delivery and parity - adjusting for age

Linear and logistic regression models – a comparison

Why do we need regression models?
 Adjustment, Effect modification, Prediction

Overview

Data to analyse	Type of analysis	Unpaired/Paired	Type	Day
Continuous	One sample mean	Irrelevant	Parametric	Day 1
			Nonparametric	Day 3
	Two sample mean	Non-paired	Parametric	Day 2
			Nonparametric	Day 2
		Paired	Parametric	Day 3
			Nonparametric	Day 3
	Regression	Non-paired	Parametric	Day 5
	Several means	Non-paired	Parametric	Day 6
			Nonparametric	Day 6
Binary	One sample mean	Irrelevant	Parametric	Day 4
	Two sample mean	Non-paired	Parametric	Day 4
		Paired	Parametric	Day 4
	Regression	**Non-paired**	**Parametric**	**Day 7**
Time to event	One sample: Cumulative risk	Irrelevant	Nonparametric	Day 8
	Regression: Rate/hazard ratio	Non-paired	Semi-parametric	Day 8

Binary regression models

There are three main types of binary regression models:
A. for the risk difference.
B. for the risk ratio.
C. for the odds ratio.

One could call them respectively binary regression models for the risk difference/risk ratio/odds ratio. Often, however, B. is called log-linear binary regression model and C. is called logistics regression model.

There is a **long tradition** for using logistic regression when considering binary outcome. Some of the reasons are:

It is the **mathematical nicest** model for binary outcome, and hence the **first** type of models that was included in the statistical software.

Binary regression models

If you have a **case-control** design, then you want to work with odds ratios.

If the **event is rare,** then it will give you **relative risk** estimates.

It is one of the few models for binary data that ensures that the estimated probability is **between zero and one**.

The logistic regression model is by far the most common model.

In many applications the logistic regression it is **not** the most natural choice, but used anyway, as in the data example today.

Binary regression models

Stata has implemented a specific function for logistic regression

```
logit outcome covariates , or
```

which is however identical to the command

```
binreg outcome covariate , or
```

Regression models for the risk ratio and risk difference can be performed by replacing the option "or" with "rr" or "rd" in the last command.

The discussion today are equally relevant for binary regression models for the risk ratio and risk difference.

Example: Post term delivery and parity

Question: How does the risk of post term delivery depend on parity?

Data: Parity and gestational age for 12,311 women in the age of 20 to 39. Post term delivery defined as a gestational age larger than 40 weeks.

Parity	N	Postterm	Risk
First child	5,938	1,722	29.0 (28.8; 30.2)%
Not first child	6,373	1,677	26.3 (25.2; 27.4)%
Total	12,311	3,399	27.6 (26.8; 28.4)%

Model: **Independent** samples from two **binomial** distributions.

Let π_0 and π_1 be the probability (risk) of post term delivery among women giving birth to their first child or not, respectively (note, π_0 and π_1 does not seem small risks).

Example: Post term delivery and parity

The assumptions behind the model was discussed on day 4.

On that day we also looked at three different measures of associations: Risk Difference, Relative Risk and Odds Ratio. And the chi-square test for no association.

Today we will look closer at the Odds Ratio.

Risk difference	-2.7 (-4.3; -1.1)%
Relative risk	0.91 (0.86; 0.96)
Odds ratio	0.87 (0.81; 0.95)

X^2=11.09 p=0.001

In the table above we compare π_1 to π_0, i.e. women giving birth to their first child is the **reference group**.

We see that the risk is (statistically significant) smaller if the woman already had a child.

Odds and risk

The odds is defined as $\pi/(1-\pi)$, i.e. the probability of post term delivery divided by the probability of not having a post term delivery.

$$odds = \frac{\pi}{1-\pi}$$

If the odds is equal to 0.5=1/2, then the risk of post term delivery is only **half** of the risk of not having a post term delivery.

We can also go from odds to risk:

$$\pi = \frac{odds}{1+odds}$$

We see that

odds = 0.5 gives π = 0.5/(1+0.5)=0.3333.

odds = 1 gives π = 1/(1+1)=0.5.

Odds and odds ratios

The odds ratio comparing parity>0 to the reference is given by

$$OR_{10} = \frac{odds_1}{odds_0} = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \frac{\pi_1 \cdot (1-\pi_0)}{\pi_0 \cdot (1-\pi_1)}$$

It is easily seen that $\pi_1 = \pi_0 \Leftrightarrow odds_1 = odds_0 \Leftrightarrow OR = 1$

OR has **nice properties**:
Switching reference group or event will just lead to 1/OR, e.g.

$$OR_{01} = \frac{odds_0}{odds_1} = \frac{1}{OR_{10}}$$

And of course the estimates and confidence intervals will transform similarly.

$$OR_{01} : \frac{1}{0.87}\left(\frac{1}{0.95};\frac{1}{0.81}\right) = 1.14(1.06;1.24)$$

Odds ratios and relative risks

The odds ratio is related to the relative risk:

$$OR_{10} = \frac{\pi_1 \cdot (1-\pi_0)}{\pi_0 \cdot (1-\pi_1)} = RR_{10} \cdot \frac{(1-\pi_0)}{(1-\pi_1)}$$

We can see that if **the event is rare**, i.e. both π_1 and π_0 are small, then the last ratio is close to 1/1=1.

So for **a rare event** we have:

$$OR \approx RR$$

Estimating the odds ratios

The odds ratio is of course estimated by:

$$\widehat{OR}_{10} = \frac{\hat{\pi}_1 \cdot (1-\hat{\pi}_0)}{\hat{\pi}_0 \cdot (1-\hat{\pi}_1)}$$

Another way to find the estimate is to make the 'classical' 2x2 table:

	Event	
Exposed	Yes	No
Yes	a	b
No	c	d

$$\widehat{OR}_{10} = \frac{a \cdot d}{b \cdot c}$$

	Post term	
Parity>0	Yes	No
Yes	1,677	4,696
No	1,722	4,216

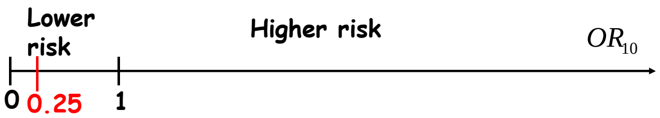
$$\widehat{OR}_{10} = \frac{1,677 \cdot 4,216}{1,722 \cdot 4,696} = 0.8743$$

Odds ratios – why inference on the log-scale

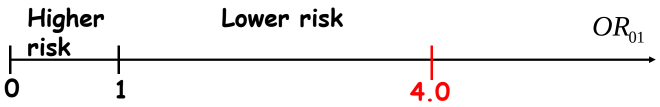
The odds ratio is limited to be positive.

A value in the interval **0 to 1** corresponds to lower risk among the Parity>0.

A value from **1 to infinity** corresponds to higher risk among the Parity>0



If we switch "exposed" and "unexposed" we get

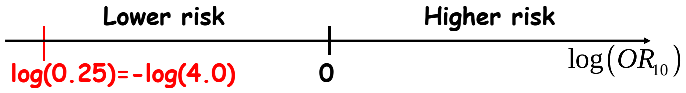


Odds ratios – why inference on the log-scale

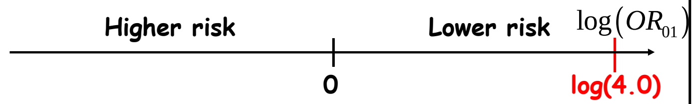
The **log-odds ratio** is not limited.

A value in the interval **-infinity to 0** corresponds to lower risk among the Parity>0.

A value from **0 to infinity** to higher risk among the Parity>0



If we switch "exposed" and "unexposed" we get



Symmetry on the log scale!!!

Odds ratios – Approx. CI (Woolf/Wald)

So **on the log scale** we have a symmetric measure of association.

On the log scale it makes sense to find the CI as 'usual', i.e. as **estimate±1.96*se** .

Using the notation from page 8 we have:

$$\text{se}\left(\ln\left(\widehat{OR}\right)\right) = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

$$\text{se}\left(\ln\left(\widehat{OR}\right)\right) = \sqrt{\frac{1}{1,677}+\frac{1}{4,696}+\frac{1}{1,711}+\frac{1}{4,216}} = 0.0403$$

$$CI\ \ln(OR) = \ln(0.8743) \pm 1.96 \cdot 0.0403 = (-0.2134; -0.0552)$$

$$CI\ OR = (\exp(-0.2134); \exp(-0.0552)) = (0.81; 0.95)$$

Odds ratios – Testing

If one wants to test a hypothesis that the odds ratio has a specific value: OR=OR_0,

then this is also done on the log-scale:

$$z_{obs} = \frac{\ln\left(\widehat{OR}\right) - \ln\left(OR_0\right)}{\text{se}\left(\ln\left(\widehat{OR}\right)\right)}$$

Could the odds be reduced by 10%, i.e. H: OR=0.9 ?

$$z_{obs} = \frac{\ln(0.8743) - \ln(0.9)}{0.0403} = -0.719$$

$$p = 2 \cdot \Pr\left(z > |z_{obs}|\right) = 2 \cdot \Pr\left(z < -|z_{obs}|\right) = 0.47$$

The hypothesis cannot be rejected.

```
disp (  ln(0.8743)-ln(0.9)  )/ 0.0403
disp 2*normal( - abs(-0.719)  )
```


Logistic regression and odds ratio

Here we will see how one can find the odds ratio by **logistic regression.**

Let ***Par1*** be an **indicator variable** for Parity>0 ,
i.e. *Par1*=1 if parity>0 and *Par1*=0 if Parity==0.

Now we will look at the (logistic regression) model:

$$\log(odds) = \beta_0 + \beta_1 \cdot Par1$$

This is equivalent to:

$$odds = \exp(\beta_0 + \beta_1 \cdot Par1) = \exp(\beta_0) \cdot \exp(\beta_1)^{Par1}$$

and

$$\pi = \frac{\exp(\beta_0 + \beta_1 \cdot Par1)}{1 + \exp(\beta_0 + \beta_1 \cdot Par1)}$$

Logistic regression and odds ratio

$\log(odds) = \beta_0 + \beta_1 \cdot Par1 \qquad odds = \exp(\beta_0) \cdot \exp(\beta_1)^{Par1}$

We see that if **Parity=0** then we have:

$\log(odds) = \beta_0 \qquad odds = \exp(\beta_0)$

and if **Parity>0** then we have

$\log(odds) = \beta_0 + \beta_1 \qquad odds = \exp(\beta_0) \cdot \exp(\beta_1)$

Combining we have

$$OR_{10} = \frac{odds(if\ parity > 0)}{odds(if\ parity = 0)} = \frac{\exp(\beta_0) \cdot \exp(\beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

Logistic regression and odds ratio

$\log(odds) = \beta_0 + \beta_1 \cdot Par1 \qquad odds = \exp(\beta_0) \cdot \exp(\beta_1)^{Par1}$

In summary we have that in the model:

The "intercept" β_0 is the **log odds in the "reference group"**.

The "slope" β_1 is the **log OR**.

That is, we can find the odds ratio from before by what is called a **logistic regression model**.

So the computer will give us **estimates** and **confidence intervals** for the **odds in the reference** group and the **odds ratio** comparing the 'exposed' to the reference.

Stata: Logistic regression and odds ratio

```
. use postterm.dta, clear
(Relationship between post term delivery, maternal age, and
parity.)
. logit ptd ib0.parity, or
Iteration 0:   log likelihood = -7253.9715
Iteration 1:   log likelihood =  -7248.429
Iteration 2:   log likelihood = -7248.4278
Iteration 3:   log likelihood = -7248.4278

Logistic regression                         Number of obs =   12311
                                            LR chi2(1)    =    1.09
                                            Prob > chi2   =  0.0009
Log likelihood = -7248.4278                 Pseudo R2     =  0.0008
----------------------------------------------------------------------
     ptd | Odds Ratio Std.Err.      z    P>|z|   [95% Conf.Interval]
-------+--------------------------------------------------------------
parity
At least one previous delivery
         |.8743241   .0352686   -3.33   0.001   .8078609   .9462552
   _cons |.4084441   .0116812  -31.31   0.000   .3861793   .4319925
----------------------------------------------------------------------
```


A simple logistic regression model

We know that the age distribution among the two groups of women is different - the women giving birth for the first time are on average younger (27.5 years versus 30.2 years in the data set).

It might be relevant to compare the two groups after "adjustment for age".

We will start by modeling the association between post term delivery and age among the **women with Parity==0**.

The simplest **logistic regression model** is:

$\log(odds) = \alpha_0 + \alpha_1 \cdot Age \qquad odds = \exp(\alpha_0) \cdot \exp(\alpha_1)^{Age}$

To get a sensible reference age:

$\log(odds) = \alpha_0 + \alpha_1 \cdot (Age - 30) \qquad odds = \exp(\alpha_0) \cdot \exp(\alpha_1)^{Age-30}$

Post term delivery and age (Parity==0)

$$\log(odds) = \alpha_0 + \alpha_1 \cdot (Age - 30) \qquad odds = \exp(\alpha_0)\cdot\exp(\alpha_1)^{Age-30}$$

$Age = 30:\ \log(odds) = \alpha_0 \qquad odds = \exp(\alpha_0)$

$Age = 31:\ \log(odds) = \alpha_0 + \alpha_1 \qquad odds = \exp(\alpha_0)\cdot\exp(\alpha_1)$

$Age = 20:\ \log(odds) = \alpha_0 - 10\cdot\alpha_1 \qquad odds = \exp(\alpha_0)\cdot\exp(\alpha_1)^{-10}$

$Age = 21:\ \log(odds) = \alpha_0 - 9\cdot\alpha_1 \qquad odds = \exp(\alpha_0)\cdot\exp(\alpha_1)^{-9}$

$Age = 25:\ \log(odds) = \alpha_0 - 5\cdot\alpha_1 \qquad odds = \exp(\alpha_0)\cdot\exp(\alpha_1)^{-5}$

$$OR_{31\text{ vs }30} = \exp(\alpha_0)\cdot\exp(\alpha_1)/\exp(\alpha_0) = \exp(\alpha_1)$$

$$OR_{21\text{ vs }20} = \exp(\alpha_0)\cdot\exp(\alpha_1)^{-9}\Big/\left(\exp(\alpha_0)\cdot\exp(\alpha_1)^{-10}\right) = \exp(\alpha_1)$$

$$OR_{25\text{ vs }20} = \exp(\alpha_0)\cdot\exp(\alpha_1)^{-5}\Big/\left(\exp(\alpha_0)\cdot\exp(\alpha_1)^{-10}\right) = \exp(\alpha_1)^5$$

Post term delivery and age (Parity==0)

$$\log(odds) = \alpha_0 + \alpha_1 \cdot (Age - 30) \qquad odds = \exp(\alpha_0)\cdot\exp(\alpha_1)^{Age-30}$$

We saw: $Age = 30:\ odds = \exp(\alpha_0)$

$$OR_{31\text{ vs }30} = \exp(\alpha_1)$$
$$OR_{21\text{ vs }20} = \exp(\alpha_1)$$
$$OR_{25\text{ vs }20} = \exp(\alpha_1)^5$$

That is,

$\exp(\alpha_0)$ is the odds in the reference (Age=30)

$\exp(\alpha_1)$ is the OR for 1 year difference.

and

$$OR_{5\,years} = OR_{1\,year}^5$$

Post term delivery and age (Parity==0)

Using a computer we get:

log odds scale	est	se	CI lower	CI upper	H = 0 z	p
Const	-0.8446	0.0332	-0.9096	-0.7795	-25.44	<0.0001
Age-30	0.0207	0.0071	0.0069	0.0345	2.93	0.003

Exp

odds scale	est	se	CI lower	CI upper	H = 1 z	p
Const	0.4297		0.4027	0.4586	-25.44	<0.0001
Age-30	1.0209		1.0069	1.0351	2.93	0.003

Note, only the estimates and the confidence intervals should be transformed!

Post term delivery and age (Parity==0)

	est	CI lower	CI upper	H = 1 z	p
Odds if Age=30	0.4297	0.4027	0.4586	-25.44	<0.0001
One years age dif.	1.0209	1.0069	1.0351	2.93	0.003

From odds to probability:

$$\Pr(\text{post term if } Age{=}{=}30) = \frac{0.4297}{1+0.4297}\left(\frac{0.4027}{1+0.4027};\frac{0.4586}{1+0.4586}\right)$$
$$= 30.1(28.7;31.4)\%$$

Five years age difference:

$$OR_{5\,years} = 1.0209^5\,(1.0069;1.0351)^5$$
$$= 1.11(1.03;1.19)$$

Stata: A simple logistic regression model

```
generate age30 = age-30
logit ptd c.age30 if parity==0, or
*** output omitted ***
Logistic regression                          Number of obs =    5938
                                               LR chi2(1)  =    8.60
                                               Prob > chi2 =  0.0034
Log likelihood = -3571.2597                    Pseudo R2   =  0.0012
---------------------------------------------------------------------
    ptd | Odds Ratio  Std. Err.     z   P>|z|   [95%Conf.Interval]
--------+------------------------------------------------------------
  age30 |   1.020915  .0072015    2.93  0.003   1.006897  1.035127
  _cons |   .4297405  .0142675  -25.44  0.000   .4026671  .4586341
---------------------------------------------------------------------
. * Odds ratio for 5 years difference
. lincom 5*age30, or
 ( 1)  5*[ptd]age30 = 0
---------------------------------------------------------------------
ptd | Odds Ratio  Std. Err.     z    P>|z|    [95% Conf. Interval]
----+----------------------------------------------------------------
(1) |   1.109039  .0391159    2.93  0.003     1.034963    1.188417
---------------------------------------------------------------------
```


Post term delivery and age (Parity==0) - Formulations

Methods
The risk of post term delivery among women giving birth for the first time was described by a logistic regression model with age as a continuous variable. …

Results
We found that five year age difference corresponds to an odds ratio of 1.11(1.03; 1.19). A 30 year old woman giving birth for the first time has 30(29;31)% risk of post term delivery.

Conclusion
The risk of post delivery among women giving birth for the first time increases with the age of the woman….

Logistic regression checking the model

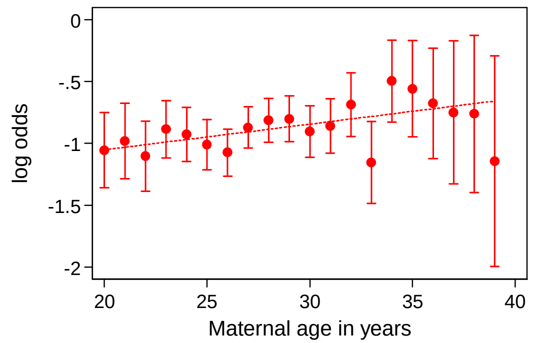
It is outside the scope of this course to go into details on how to check the model, so we will just state the assumptions behind the model:

1. All the observations should be **independent**.
2. There is exactly the same **two possible outcomes** for each observation.
3. The log odds is a **linear function** of age.

The last assumption can to some extend be checked by plotting the fitted regression line and the observed odds (with 95% CI) for each distinct age.

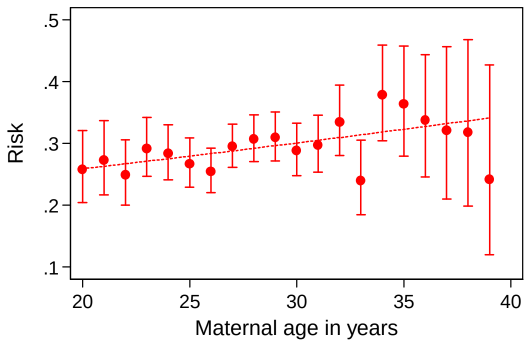
Post term delivery and age (Parity==0)

$$\log(odds) = \alpha_0 + \alpha_1 \cdot (Age - 30)$$



Post term delivery and age (Parity==0)

$$risk = \frac{\exp\left(\alpha_0 + \alpha_1 \cdot (Age - 30)\right)}{1 + \exp\left(\alpha_0 + \alpha_1 \cdot (Age - 30)\right)}$$



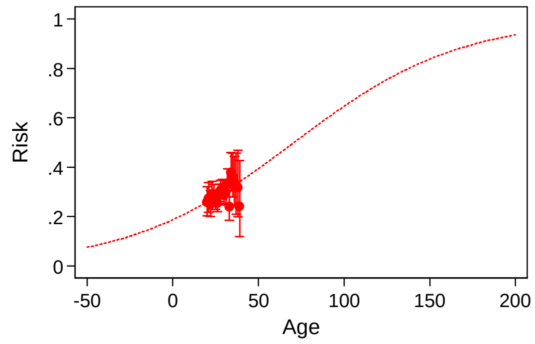
Not linear! But almost

Erik Parner Basic Biostatistics - Day 7 29

Post term delivery and age (Parity==0)

Extrapolating we see the non-linearity.

$$risk = \frac{\exp\left(\alpha_0 + \alpha_1 \cdot (Age - 30)\right)}{1 + \exp\left(\alpha_0 + \alpha_1 \cdot (Age - 30)\right)}$$



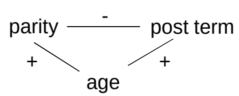
Erik Parner Basic Biostatistics - Day 7 30

Post term delivery parity adjusting for age

We now know that

- The risk of post term delivery **increases with age** (among women with parity==0).
- The risk of post term delivery **is smaller for Parity>0**.
- Women with Parity>0 are older.

From this we can deduce that **adjusting for age** (if reasonable) will **increase** the difference between the two parity groups.



- negative association
+ positive association

We now show how to find an age adjusted estimate, when we assume a linear "effect" of age on log odds.

Erik Parner Basic Biostatistics - Day 7 31

Post term delivery and age

We fit the same model to the Parity>0 group and then look at the difference:

log odds	Slope				log odds Age==30			
	est	se	lower	upper	est	se	lower	upper
Parity>0	0.025	0.007	0.012	0.039	-1.037	0.029	-1.093	-0.981
Parity==0	0.021	0.007	0.007	0.035	-0.845	0.033	-0.910	-0.780
Difference	0.005	0.010	-0.015	0.024	-0.193	0.044	-0.279	-0.107

The standard errors of the differences are found as usual:

$$\mathrm{se}\left(est_{parity>0} - est_{parity==0}\right) = \sqrt{\mathrm{se}^2_{parity>0} + \mathrm{se}^2_{parity==0}}$$

We see that we can assume the slopes to be identical, we could also test the hypothesis:

$$z = \frac{0.005 - 0}{0.010} = 0.5 \qquad p = 64\%$$

Erik Parner Basic Biostatistics - Day 7 32

Post term delivery and age - assuming identical slopes

If we **assume identical slopes**, then we can write the model:

$$\log(odds) = \gamma_0 + \gamma_1 \cdot (Age - 30) + \gamma_2 \cdot Par1$$

$$odds = \exp(\gamma_0) \cdot \exp(\gamma_1)^{(Age-30)} \cdot \exp(\gamma_2)^{Par1}$$

We see that

$\exp(\gamma_0)$ is the odds among 30-year old with Parity==0.

$\exp(\gamma_0) \cdot \exp(\gamma_2)$ is the odds among 30-year old with Parity>0.

$\exp(\gamma_1)^A$ is the odds ratio for the age difference A years among women in the same Parity group.

$\exp(\gamma_2)$ is the odds ratio comparing Parity>0 to Parity==0 , at the same age.

Post term delivery and age - assuming identical slopes

The model is easily fitted by a computer:

log odds	Slope est	Slope se	Slope lower	Slope upper	log odds Age==30 est	se	lower	upper
Parity>0	0.023	0.005	0.013	0.033	-1.036	0.029	-1.092	-0.981
Parity==0					-0.839	0.031	-0.900	-0.778
Difference	0				-0.197	0.043	-0.281	-0.114

The **age adjusted** OR comparing Parity>0 to Parity==0:

$$OR_{10} : \exp(-0.197)(\exp(-0.281);\exp(-0.114)) = 0.821(0.755;0.892)$$

and if we compare Parity==0 to Parity>0:

$$OR_{01} : \frac{1}{0.821}\left(\frac{1}{0.892};\frac{1}{0.755}\right) = 1.22(1.12;1.32)$$

Stata: Two simple logistic regression models

A logistic regression with the β coefficients in the output:

```
. * The two models in the same analysis, such that
. * we can compare the two groups
. logit ptd ib0.parity##c.age30
  *** output omitted ***
------------------------------------------------------------------------------
         ptd |      Coef.  Std. Err.      z    P>|z|    [95% Conf.Interval]
-------------+----------------------------------------------------------------
      parity |
At least one previous delivery
             |  -.1927641   .0438288   -4.40   0.000    -.278667  -.1068611
       age30 |   .0206988    .007054    2.93   0.003    .0068732   .0345244
             |
parity#c.age30 |
At least one previous delivery
             |   .0045463   .0098178    0.46   0.643   -.0146963   .0237888
             |
       _cons |  -.8445738   .0332003  -25.44   0.000   -.9096451  -.7795025
------------------------------------------------------------------------------
```


Stata: Two simple logistic regression models

b0: parity=0 is set to be the ref.

##: we allow for different slopes

c: age30 is considered continuous with linear effect

```
 logit ptd ib0.parity##c.age30
  *** output omitted ***
------------------------------------------------------------------
         ptd |      Coef.  Std. Err.      z    P>|z|    [95
-------------+----------------------------------------------------
      parity |
At least one previous delivery
             |  -.1927641   .0438288   -4.40   0.000    -.
       age30 |   .0206988    .007054    2.93   0.003    .0
             |
parity#c.age30 |
At least one previous delivery
             |   .0045463   .0098178    0.46   0.643   -.0
             |
       _cons |  -.8445738   .0332003  -25.44   0.000   -.9
------------------------------------------------------------------
```

Difference in intercept for parity 1 and 0 (parity)

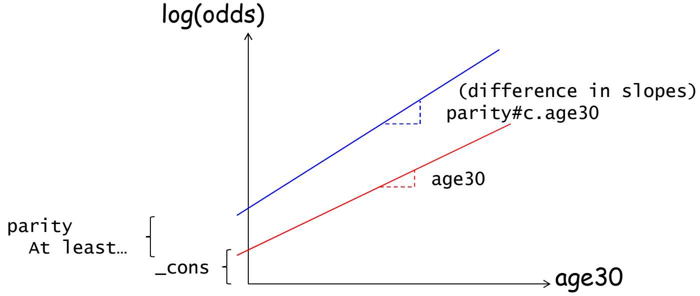
Slope for parity 0 (age30)

Difference in slope for parity 1 and 0 (parity#c.age30)

Intersection for parity 0 (_cons)

Stata: Two simple logistic regression models

The estimates can be placed on a graph



Stata: Two simple logistic regression models

```
. * The two models in the same analysis, such that
. * we can compare the two groups
. logit ptd ib0.parity c.age30
  *** output omitted ***
------------------------------------------------------------------------------
          ptd |      Coef.   Std. Err.      z    P>|z|   [95% Conf.Interval]
--------------+---------------------------------------------------------------
       parity |
At least one previous delivery
              | -.1974712   .0425807   -4.64   0.000   -.2809278 -.1140145
        age30 |  .0230464   .0049045    4.70   0.000    .0134337   .032659
        _cons | -.8389922   .0308945  -27.16   0.000   -.8995443 -.7784401
------------------------------------------------------------------------------
. logit, or
------------------------------------------------------------------------------
          ptd | Odds Ratio Std. Err.      z    P>|z|   [95% Conf.Interval]
--------------+---------------------------------------------------------------
       parity |
At least one previous delivery
              |  .8208038   .0349504   -4.64   0.000    .7550829    .892245
        age30 |  1.023314   .0050188    4.70   0.000    1.013524   1.033198
        _cons |  .4321458   .0133509  -27.16   0.000     .406755   .4591216
------------------------------------------------------------------------------
```


Stata: Two simple logistic regression models

```
. logit ptd ib0.parity c.age30
  *** output omitted ***
------------------------------------------------------------
          ptd |      Coef.   Std. Err.      z    P>|z|   [95
--------------+---------------------------------------------
       parity |
At least one previous delivery
              | -.1974712   .0425807   -4.64   0.000   -.2
        age30 |  .0230464   .0049045    4.70   0.000    .0
        _cons | -.8389922   .0308945  -27.16   0.000   -.8
------------------------------------------------------------
```

Difference in intercept for parity 1 and 0 (→ parity)

Slope for both parity 0 and 1 (→ age30)

Intersection for parity 0 (→ _cons)

Post term delivery and Parity - Formulations

Methods
The risk of post term delivery among women was modeled by a logistic regression model with age as a continuous variable....

Results version1
Comparing Parity >0 to Parity==0 the crude odds ratio was 0.87(0.81;0.95). The age adjusted odds ratio was 0.82(0.76;0.89).

Results version2
Comparing Parity==0 to Parity>0 the crude odds ratio was 1.14(1.06;1.24). The age adjusted odds ratio was 1.22(1.12;1.32).

Conclusion ??
Women giving birth for the first time have a 22% (12%;32%) higher **odds** of post term delivery compared to other women of the same age. .....

The odds can be predicted by

$$odds = \exp(\gamma_0 + \gamma_1 \cdot (Age - 30) + \gamma_2 \cdot Par1)$$

The risk can be predicted by

$$risk = \frac{\exp(\gamma_0 + \gamma_1 \cdot (Age - 30) + \gamma_2 \cdot Par1)}{1 + \exp(\gamma_0 + \gamma_1 \cdot (Age - 30) + \gamma_2 \cdot Par1)}$$

Odds ratio is not always relative risk

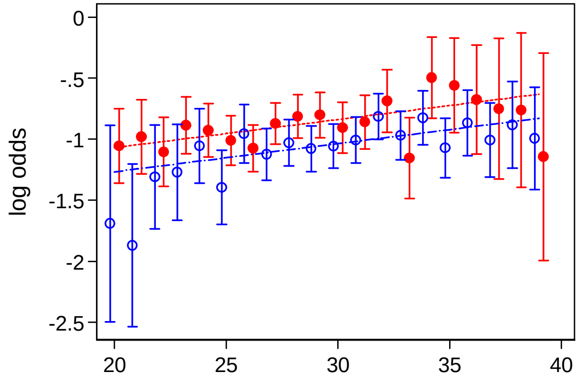
A women of age 30 in the data set:

Parity 0:	odds=0.3547	risk=0.2618
Parity 1:	odds=0.4321	risk=0.3017
	OR=1.22	RR=1.15

Erik Parner Basic Biostatistics - Day 7 41

Post term delivery and age - assuming identical slopes

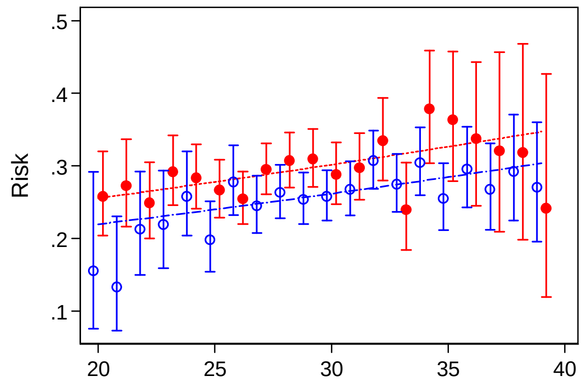
$$\log(odds) = \gamma_0 + \gamma_1 \cdot (Age - 30) + \gamma_2 \cdot Par1$$



Erik Parner Basic Biostatistics - Day 7 42

Post term delivery and age - assuming identical slopes

$$risk = \frac{\exp(\gamma_0 + \gamma_1 \cdot (Age - 30) + \gamma_2 \cdot Par1)}{1 + \exp(\gamma_0 + \gamma_1 \cdot (Age - 30) + \gamma_2 \cdot Par1)}$$



Erik Parner Basic Biostatistics - Day 7 43

Linear and logistic regression - a comparison

In a **linear regression** the **outcome** is **continuous**:
Lung function, Blood pressure, BMI, concentrations…

In a **logistic regression** the **outcome** is **binary**:
Post term delivery, gender, dead/alive, sick/ well, BMI>30.

Neither of the models make **any assumptions** about the **explanatory variable**!!
In both models they can be continuous, binary or categorical.

In **both models** we have to assume **independence** between observations.

In **both models** we assume linearity –
of expected value or the log odds.

Both models are readily fitted by standard statistical packages.

Erik Parner Basic Biostatistics - Day 7 44

Regression models in general – why? Adjustment

You have now looked at two of the most commonly used regression models in their most simple forms, involving one continuous and one binary explanatory variable.

You have seen how one can use such models for **adjustment**: What is the 'effect' of the binary 'exposure' when **adjusting** for the continuous variable?

Exactly the same models could answer the question: What is the 'effect' (slope) of the **continuous** 'exposure' when **adjusting** for the binary variable?
E.g. what is the increase in risk of post term delivery associated with age when we **adjust** for parity?

Often one has several explanatory variables, a mixture of continuous, binary and categorical and the purpose is to **adjust for more than one**.
In such case one might apply a **multiple regression** model.

Regression models in general – why? Effect modification

We have also seen how we could compare the 'effect' of one explanatory variable for subgroups described by another explanatory variable (**effect modification**):

What is difference in the PEFR-height relationship for men and women?
What is the difference in the Risk-age relationship for the two parity groups?

Typically by **comparing the slopes**.

Often one has several explanatory variables, a mixture of continuous, binary and categorical and the purpose is to **model effect modification** between explanatory variables.
In such case one might apply a **multiple regression** model.

Regression models in general – why? Prediction

We could also have used exactly the same models for **prediction/prognosis**:

What is the expected PEFR for a person with a given sex and a given height?
What is the risk of post term delivery for women of a given age having her first child?

Often one has several explanatory variables, a mixture of continuous, binary and categorical and the purpose is to make **prediction** for a person with a given set of characteristics.
In such a case one might apply a **multiple regression** model.