Erik Parner
27-10-2015

PhD Course in Basic Biostatistics

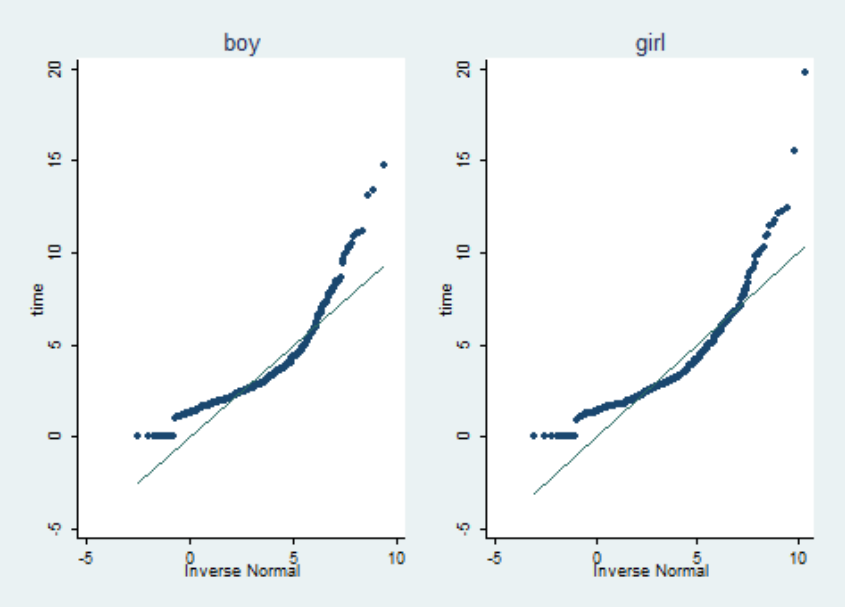
Fall 2015

1. Consider the chance of getting a boy the second time. Estimate this if the older sib was a boy and if she was a girl.
   Test the hypothesis of no association between the sex of the first and second child.

Data are analyzed as two independent samples from two binomial distributions. A chi-squared test is used to assess the hypothesis of no association. The chance of getting a boy when the older sibling is a boy is 50.1% (95% CI: 45.7%-54.5%) compared to 53.7% (95% CI: 49.1%-58.2%) when the older sibling is a girl. This reduction of 3.6%-points (95% CI: -2.6%-points-9.8%-points) is not statistically significant (p=0.26). Similarly, the statistically non-significant relative risk for getting a boy is 0.93 (95% CI: 0.83-1.05).

2. Is the time between births of the two children depending on the sex of the first born?

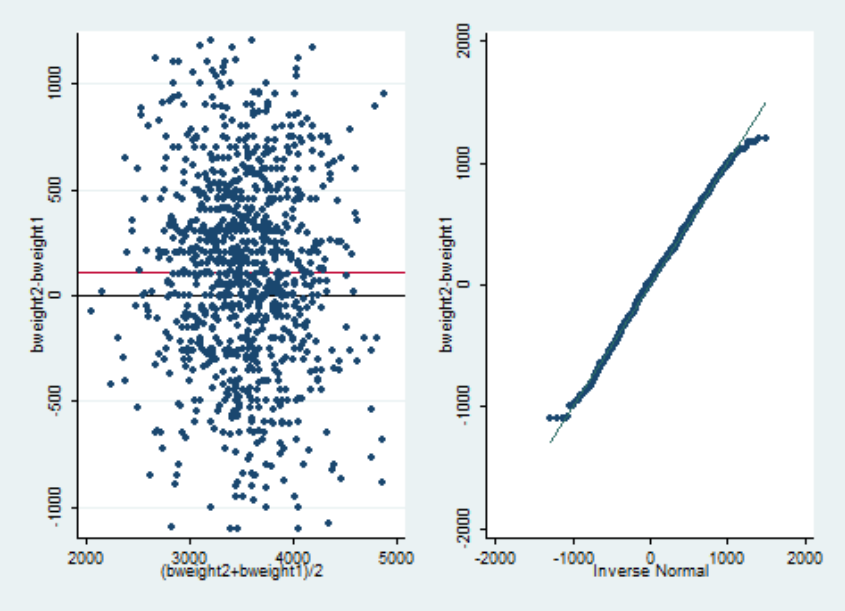
Judging from the QQ-plots the time between births of the two children (mage2-mage1) cannot be assumed to follow a normal distribution for any of the sexes of the first born child.



The median time between births of two children born to women whose first born is a boy is 2.88 (95% CI: 2.79-3.00) while the median time between births of two children born to women whose first born is a girl is 3.02 (95% CI: 2.86-3.15). The Wilcoxon rank sum test is used to compare the time between births of two children born to women whose first born child is a boy to women whose first born child is a girl. Time between births of two children is not statistically significantly different when comparing boys and girls (p=0.29).

3. Compare the birth weight of the first and second born.
   Compare the chance of a birth weight below 3600g of the first and second born.

Data are analyzed as a paired sample based on Student's *t*-test for paired samples. The assumptions are checked by a Bland-Altman plot and a QQ-plot of the differences. Judging from the Bland-Altman plot it seems that the mean level and variation of the absolute difference is approximately the same for all women, hence the difference (bweight2-bweight1) is considered an appropriate measure for describing the change in birth weight.



The increase in birth weight from first to second born child was 109 (SD=452) g. The mean increase of 109 (95% CI: 81-137) g was statistically significant (p<0.0001).

The difference in incidence of birth weight below 3600 g for second born children compared to first born children was described by a risk difference. The hypothesis of no difference in risk was tested by McNemar's test. The incidence of birth weight below 3600 g was 51.2% (95% CI: 48.1%-54.3%) for second born children and 58% (95% CI: 54.9%-61.1%) for first born children, corresponding to a reduction in incidence of 6.8%-points (95 CI: 3.3%-points-10.3%-points). The difference is statistically significant (p=0.0001).

4. Predict the birth weight of the second born (hint: create a prediction interval).
   Predict the birth weight of the second born using the sex of the second born.

The prediction interval for birth weight of the second born child is (2602-4560) g, that is, approximately 95% of the children in a comparable population (of second born children) will have a birth weight between 2602 and 4560 g. 2.5% will be below 2602 g and 2.5% will be above 4560 g. For boys the prediction interval is (2667-4655) g while for girls the prediction interval is (2561-4429) g.

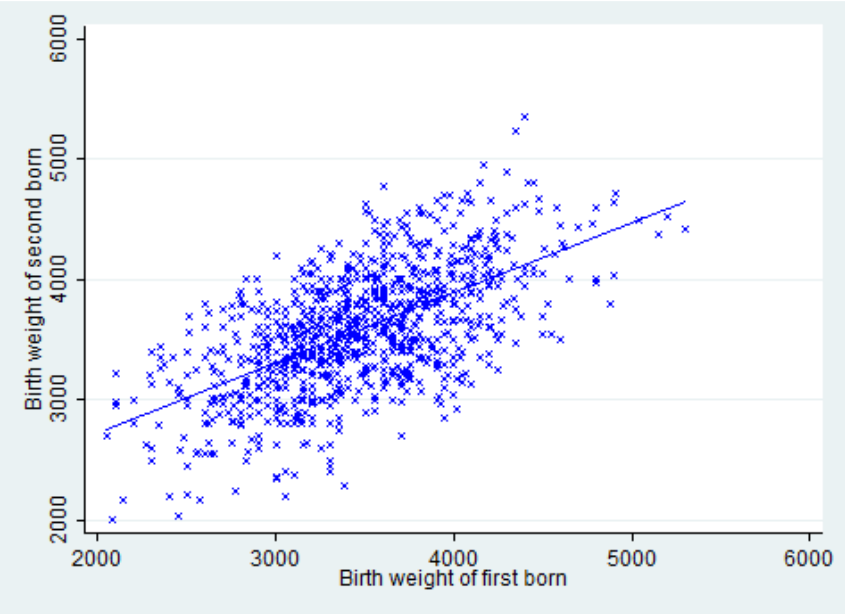
5. Predict the birth weight of the second born if the birth weight of the first born is 3000g (hint: establish a linear regression with the birth weight of the second born as outcome and birth weight of the first born as explanatory variable).

Present a prediction formula for the birth weight of the second born for various values of birth weights of the first born.

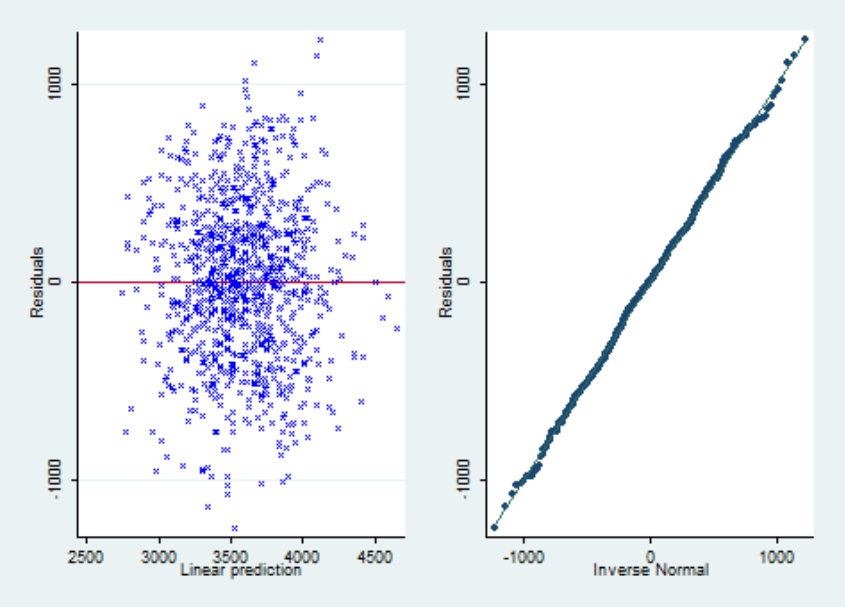
Let bweight2$_i$ and bweight1$_i$ be the birth weight for the second born child and the first born child, respectively. The model

$$\text{bweight2}_i = \beta_0 + \beta_1 \cdot \text{bweight1}_i + \text{E}_i \qquad \text{E}_i \sim N(0, \sigma^2)$$

seems appropriate considering the scatter plot.



The model is checked by diagnostic plots of the residuals, which look nice.



The model

$$\text{bweight2}_i = \beta_0 + \beta_1 \cdot \text{bweight1}_i + \text{E}_i \qquad \text{E}_i \sim N(0, \sigma^2)$$

is fitted, resulting in estimates of β_0, β_1 and σ:

$\hat{\beta}_0 =$ 1551.63 (95% CI: 1384.68-1718.58) g

$\hat{\beta}_1 =$ 0.58 (95% CI: 0.54-0.63) g/g

$\hat{\sigma} =$ 397.18 g

Particularly, a difference in birth weight of one gram for first born children corresponds to a difference in mean birth weight of 0.58 (95% CI: 0.54-0.63) g for second born children.

Based on the estimates, a prediction formula for birth weight of the second born for various values of birth weight of the first born can be expressed as

$$\widehat{\text{bweight2}}_i = 1551.63 + 0.58 \cdot \text{bweight1}_i$$

In particular, the predicted value of birth weight of the second born if the first born is 3000 g is 3305 (95% CI: 3272-3339) g with corresponding 95% PI of (2527-4084) g.

6. Establish a multiple linear regression model with the birth weight of the second born as outcome and
   A. birth weight of the first born
   B. sex of the second born

   as explanatory variables.
   Test the hypothesis of equal slope for boys and girls.
   Based on the linear regression model predict the birth weight of the second born boy if the birth weight of the first born is 3000g.
   Present a prediction formula for the birth weight of the second born for various values of birth weights of the first born and sex of the second born.

The difference in slope between girls and boys is -0.05 (95% CI: -0.15-0.04), which is not statistically significant (p=0.28). The model is checked by diagnostic plots of the residuals within boys and girls, respectively, which look nice.

The model, when defining girl2 to be the binary variable for girl, is given by

$$\text{bweight2}_i = \beta_0 + \beta_1 \cdot \text{bweight1}_i + \beta_2 \cdot \text{girl2}_i + \text{E}_i \qquad \text{E}_i \sim N(0, \sigma^2)$$

which fitted gives rise to the estimates of β_0, β_1, β_2 and σ:

$\hat{\beta}_0 =$ 1639.95 (95% CI: 1472.59-1807.32)

$\hat{\beta}_1 =$ 0.58 (95% CI: 0.53-0.63)

$\hat{\beta}_2 =$ -138.66 (95% CI: -187.31 - -90.01)

$$\hat{\sigma} = 391.29$$

Based on the estimates, a prediction formula for birth weight of the second born for various values of birth weight of the first born and the sex of the second born can be expressed as

$$\widehat{\text{bweight}}2_i = 1639.95 + 0.58 \cdot \text{bweight}1_i - 138.66 \cdot \text{girl}2_i$$

In particular, the predicted value of birth weight of the second born boy if birth weight of the first born is 3000 g is 3375 (95% CI: 3334-3416) g with corresponding 95% PI of (2608-4142) g.

Do-file

```
**********************************************
* Solution.
**********************************************

cd "U:\Teaching\BasalBiostat\Exam new\"
capture log close
log using "Solution.txt", replace text

use bweight, clear


**********************************************
* Question 1.
**********************************************
generate boy1=(sex1==1)
generate boy2=(sex2==1)
label define NoYes 0 "No" 1 "Yes"
label val boy1 NoYes
label val boy2 NoYes
tabu boy1 sex1
tabu boy2 sex2

tabu boy1 boy2, row
ci boy2 if boy1==0, bin
ci boy2 if boy1==1, bin
cs boy2 boy1, or


**********************************************
* Question 2.
**********************************************
gen time=mage2-mage1
qnorm time if(boy1==1), title(boy)
qnorm time if(boy1==0), title(girl)
centile time if(boy1==1), c(50)
centile time if(boy1==0), c(50)
ranksum time, by(boy1)


**********************************************
* Question 3.
**********************************************
gen diff=bweight2-bweight1
gen ave=(bweight2+bweight1)/2
label variable diff "bweight2-bweight1"
label variable ave "(bweight2+bweight1)/2"

sum diff
```

```
local diffmean=r(mean)
scatter diff ave, yline(0, lco(black)) yline(`diffmean')
qnorm diff

ttest bweight2=bweight1

gen below1=(bweight1<3600)
gen below2=(bweight2<3600)
ci below1, bin
ci below2, bin
mcc below2 below1


**********************************************
* Question 4.
**********************************************
centile bweight2, c(2.5 97.5) meansd
centile bweight2 if(sex2==1), c(2.5 97.5) meansd
centile bweight2 if(sex2==2), c(2.5 97.5) meansd


**********************************************
* Question 5.
**********************************************
regress bweight2 bweight1
twoway ///
                (scatter bweight2 bweight1, mco(blue) msy(x)) ///
                (lfit bweight2 bweight1 , lco(blue)) ///
, ytitle("Birth weight of second born") legend(off)
predict fit if e(sample), xb
predict res if e(sample), res
scatter res fit, yline(0) mco(blue) msy(x)
qnorm res
drop fit res

regress bweight2 bweight1
lincom _cons+3000*bweight1
disp r(estimate)-1.96*e(rmse), r(estimate)+1.96*e(rmse)


**********************************************
* Question 6.
**********************************************
regress bweight2 bweight1 if(sex2==1)
twoway ///
                (scatter bweight2 bweight1 if(sex2==1), mco(blue) msy(x)) ///
                (lfit bweight2 bweight1 if(sex2==1), lco(blue)) ///
, ytitle("Birth weight of second born") legend(off)
predict fitB if e(sample), xb
predict resB if e(sample), res
```

```
scatter resB fitB, yline(0) mco(blue) msy(x)

regress bweight2 bweight1 if(sex2==2)
twoway ///
                (scatter bweight2 bweight1 if(sex2==2), mco(blue) msy(x)) ///
                (lfit bweight2 bweight1 if(sex2==2), lco(blue)) ///
, ytitle("Birth weight of second born") legend(off)
predict fitG if e(sample), xb
predict resG if e(sample), res
scatter resG fitG, yline(0) mco(blue) msy(x)

regress bweight2 b1.sex2##c.bweight1

regress bweight2 b1.sex2 c.bweight1

lincom _cons+3000*bweight1
disp r(estimate)-1.96*e(rmse), r(estimate)+1.96*e(rmse)


log close
```