folkehelseinstituttet

Stata:
Linear Regression

Stata 3, linear regression

Hein Stigum

Presentation, data and programs at:

http://folk.uio.no/heins/

courses

folkehelseinstituttet

Birth weight by gestational age

SYNTHETIC DATA EXAMPLE

Linear regression

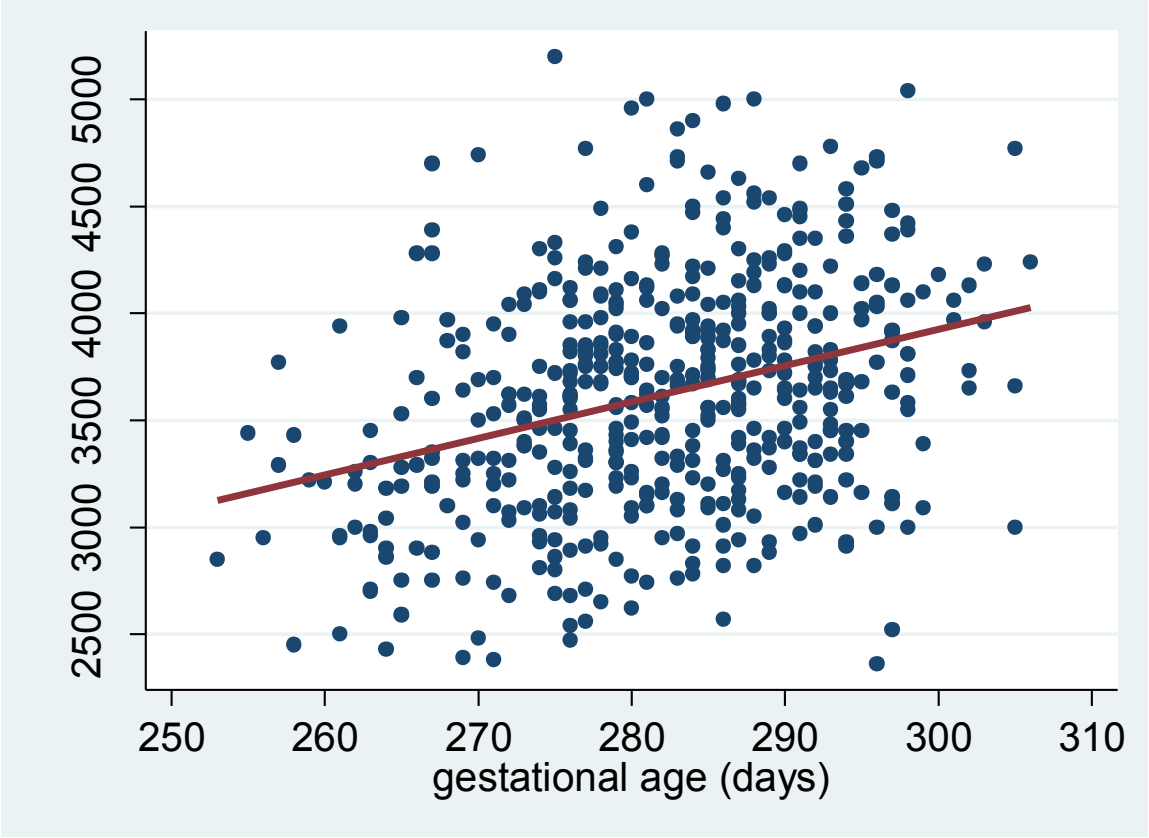
Birth weight

by

gestational age

folkehelseinstituttet

Regression idea



$$\text{model}: \quad y = b_0 + b_1 x + e$$

$$\text{y} = \text{outcome}$$

$$\text{x} = \text{covariate}$$

$$\text{b}_1 = \text{coefficient, effect of x}$$

$$\text{e} = \text{error, residual}$$

$$\text{model with many cofactors}: \quad y = b_0 + b_1 x_1 + b_2 x_2 + e$$

$$\text{x}_1 \text{ , x}_2 = \text{covariate}$$

Model, measure and assumptions

- Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \quad \varepsilon \propto N(0, \sigma^2)$$

- Association measure


$_1$ = change in y for one unit increase in x_1

- Assumptions
  - Independent errors
  - Linear effects
  - Constant error variance

- Robustness
  - influence

folkehelseinstituttet

Association measure

Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Start with:

$$\Delta y_1 = y_{x_1=2} - y_{x_1=1}$$
$$= \beta_0 + 2\beta_1 + \beta_2 x_2$$
$$- \beta_0 - 1\,\beta_1 - \beta_2 x_2$$
$$= \beta_1$$

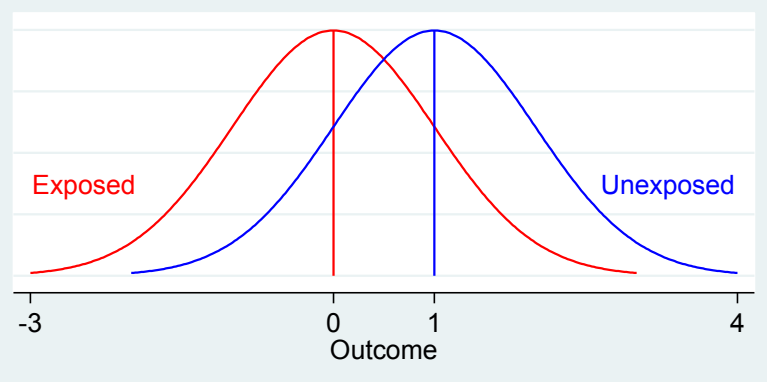
Hence: $\Delta y_1 = \beta_1$

Purpose of regression

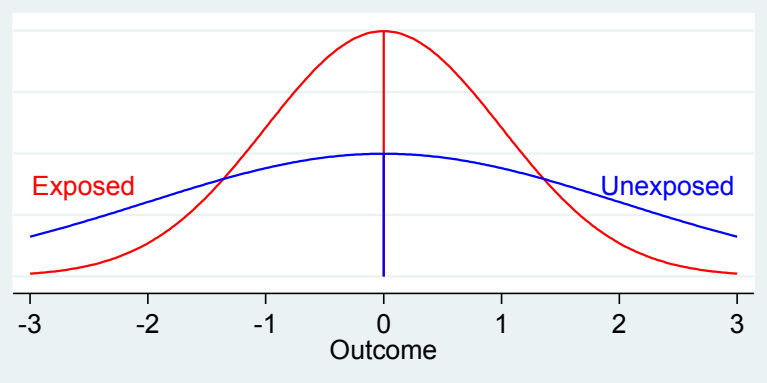
- Estimation
  - Estimate association between outcome and exposure adjusted for other covariates
- Prediction
  - Use an estimated model to predict the outcome given covariates in a new dataset

folkehelseinstituttet

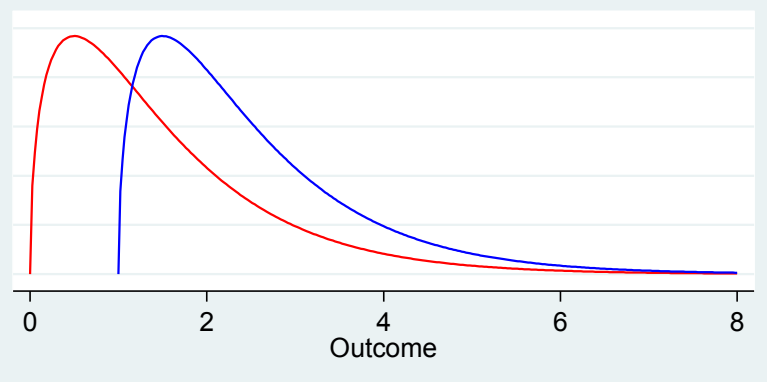
Outcome distributions by exposure



Linear regression



Quantile regression
or
cutoff,
logistic regression

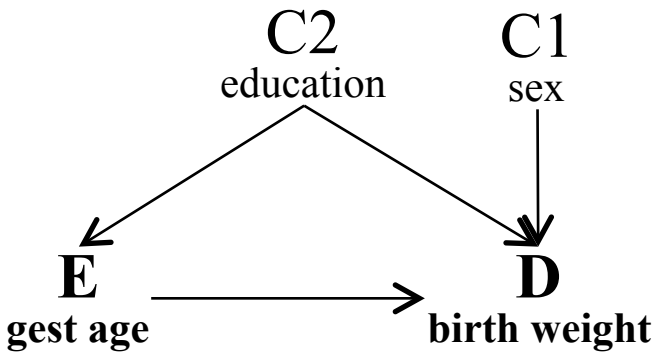


Linear regression
or
transform,
linear regression

folkehelseinstituttet

Workflow

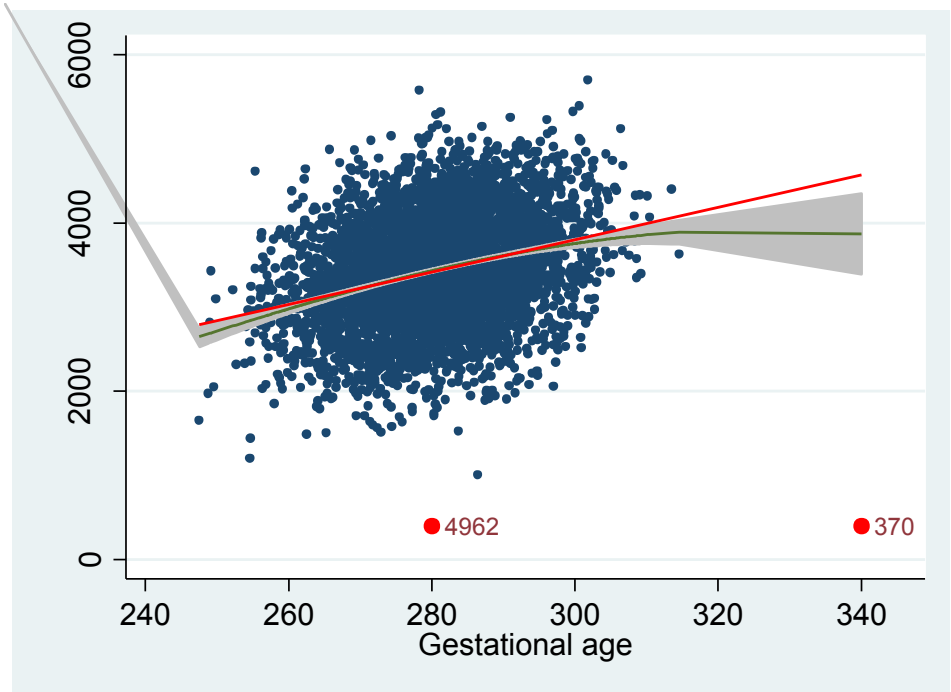
- DAG
- Scatter- and densityplots
- Bivariate analysis
- Regression
  - Model estimation
  - Test of assumptions
    - Independent errors
    - Linear effects
    - Constant error variance
  - Robustness
    - Influence



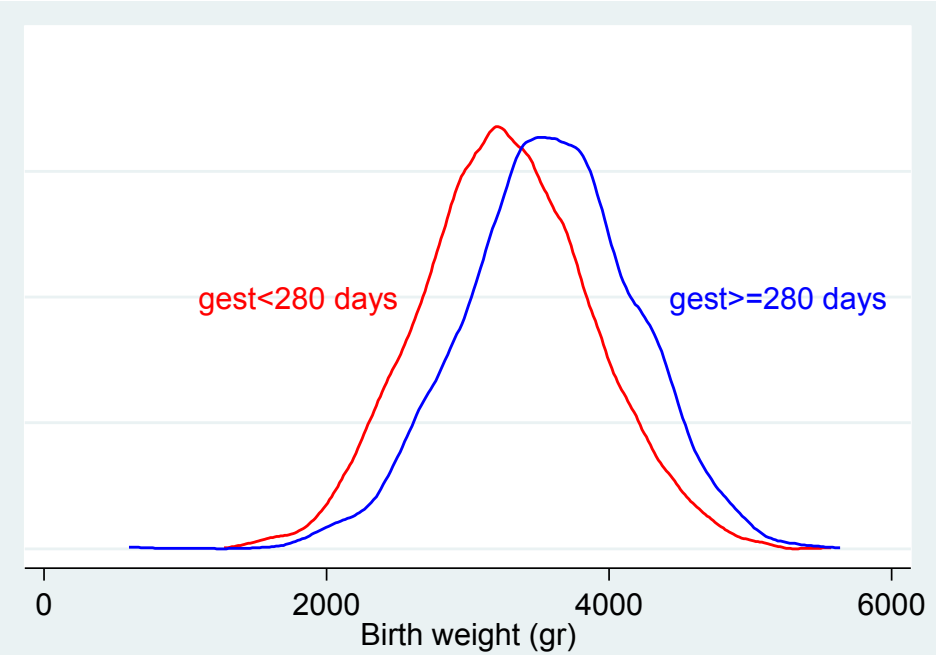
folkehelseinstituttet

Scatter and density plots

Scatter of birth weight by gestational age



Distribution of birth weight for low/high gestational age



Look for deviations from linearity and outliers

Look for shift in shape

Syntax

- Estimation
  - regress y x1 x2 — linear regression
  - regress y c.age i.sex — continuous age, categorical sex
  - regress y c.age##i.sex — main+interaction
- Compare models
  - estimates store m1 — save model
  - estimates table m1 m2 — compare coefficients
  - estimates stats m1 m2 — compare model fit
- Post estimation
  - predict res, residuals — predict residuals

folkehelseinstituttet

Model 1: outcome+exposure

regress bw gest

crude model

Source	SS	df	MS
Model	177917413	1	177917413
Residual	1.7861e+09	4998	357360.133
Total	1.9640e+09	4999	392879.247

Number of obs	5000
F(1, 4998)	497.87
Prob > F	0.0000
R-squared	0.0906
Adj R-squared	0.0904
Root MSE	597.8

bw	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
gest	19.2421	.8623753	22.31	0.000	17.55147	20.93274
_cons	-1971.168	242.3085	-8.13	0.000	-2446.199	-1496.137

estimates store m1

store model results

folkehelseinstituttet

Model 2 and 3: Add covariates

regress bw gest i.educ sex

add covariates

estimates table m1 m2 m3

compare coefs

Variable	m1	m2	m3
gest	19.2	17.9	17.9
educ			
2		70.6	71.5
3		99.0	99.1
sex			-154.3
_cons	-1971.2	-1652.1	-1572.3

Estimate association:
m1 is biased, m2=m3

m3 more precise?

estimates stats m1 m2 m3

compare fit

Model	Obs	ll(null)	ll(model)	df	AIC
m1	5000	-39297.34	-39059.94	2	78123.88
m2	5000	-39297.34	-39051.98	4	78111.96
m3	5000	-39297.34	-39009.84	5	78029.69

Prediction:
m3 is best

folkehelseinstituttet

Factor (categorical) variables

- Variable
  - educ = 1, 2, 3 for low, medium and high education
- Built in
  - i.educ use educ=1 as base (reference)
  - ib3.educ use educ=3 as base (reference)
- Manual “dummies”
  - educ=1 as base, make dummies for 2 and 3
  - generate Medium =(educ==2) if educ<.
  - generate High =(educ==3) if educ<.

folkehelseinstituttet

Create meaningful constant

Expected birth weight:

$$E(bw) = \beta_0 + \beta_1 \cdot gest + \beta_2 \cdot educ2 + \beta_3 \cdot educ3 + \beta_4 \cdot sex$$

Expected birth weight at:

gest= 0, educ=1, sex=0, not meaningful $\quad \hat{\beta}_0 = -1572gr$

gest=280, educ=1, sex=0 $\quad \hat{\beta}_0 + \hat{\beta}_1 \cdot 280 = 3426gr$

Margins:

margins, at(gest= 0 educ=1 sex=0) = -1572 not meaningful

margins, at(gest= 280 educ=1 sex=0) = 3426

folkehelseinstituttet

Results so far

	coeff	95% conf. Int.
Birth weight at ref	3426	(3385 , 3467)
Gestational age		
per day	17.9	(16 , 20)
Education		
Low	0	
Medium	71.5	(25 , 118)
High	99.1	(51 , 148)
Sex		
Boy	0	
Girl	-154.3	(-187 , -121)

Would normally check for interaction now!

folkehelseinstituttet

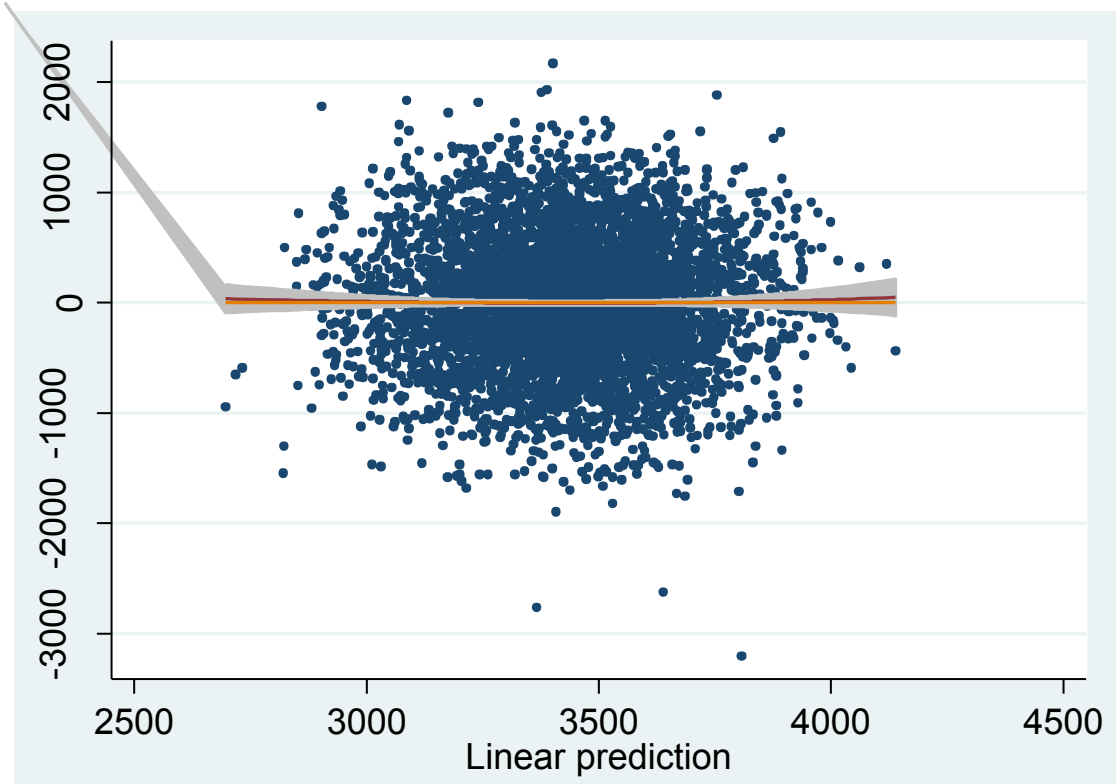
ASSUMPTIONS

folkehelseinstituttet

Test of assumptions

- Assumptions
  - Independent residuals: discuss
  - Linear effects: } plot residuals versus predicted y
  - Constant variance: }

predict res, residuals
predict pred, xb
scatter res pred

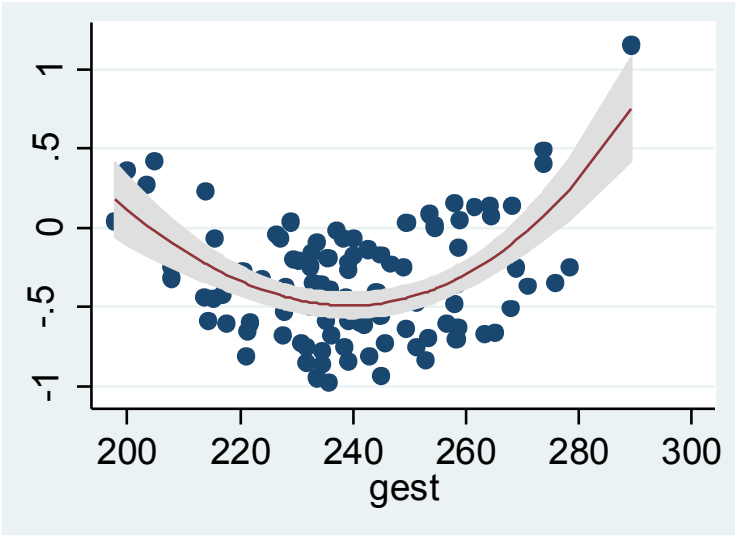


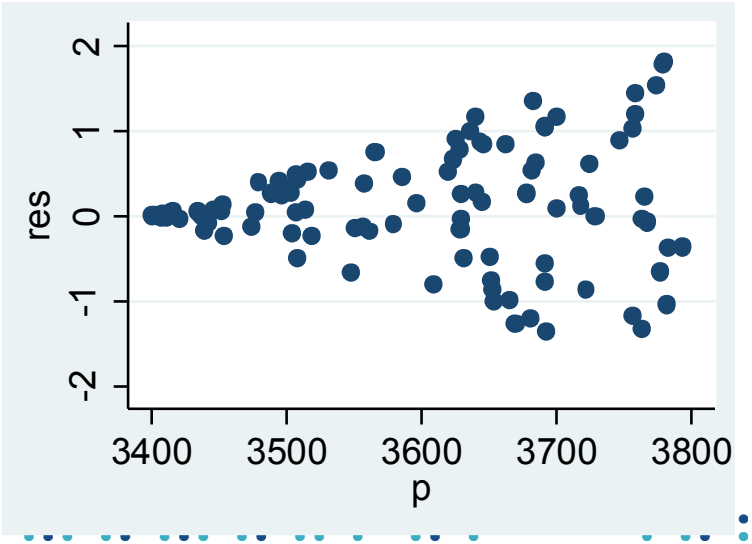
estat hettest
p=0.9
no heteroskedasticity

folkehelseinstituttet

Violations of assumptions

- Dependent residuals
  - Use mixed models or GEE
- Non linear effects
  - Add square term or spline
- Non-constant variance
  - Use robust variance estimation





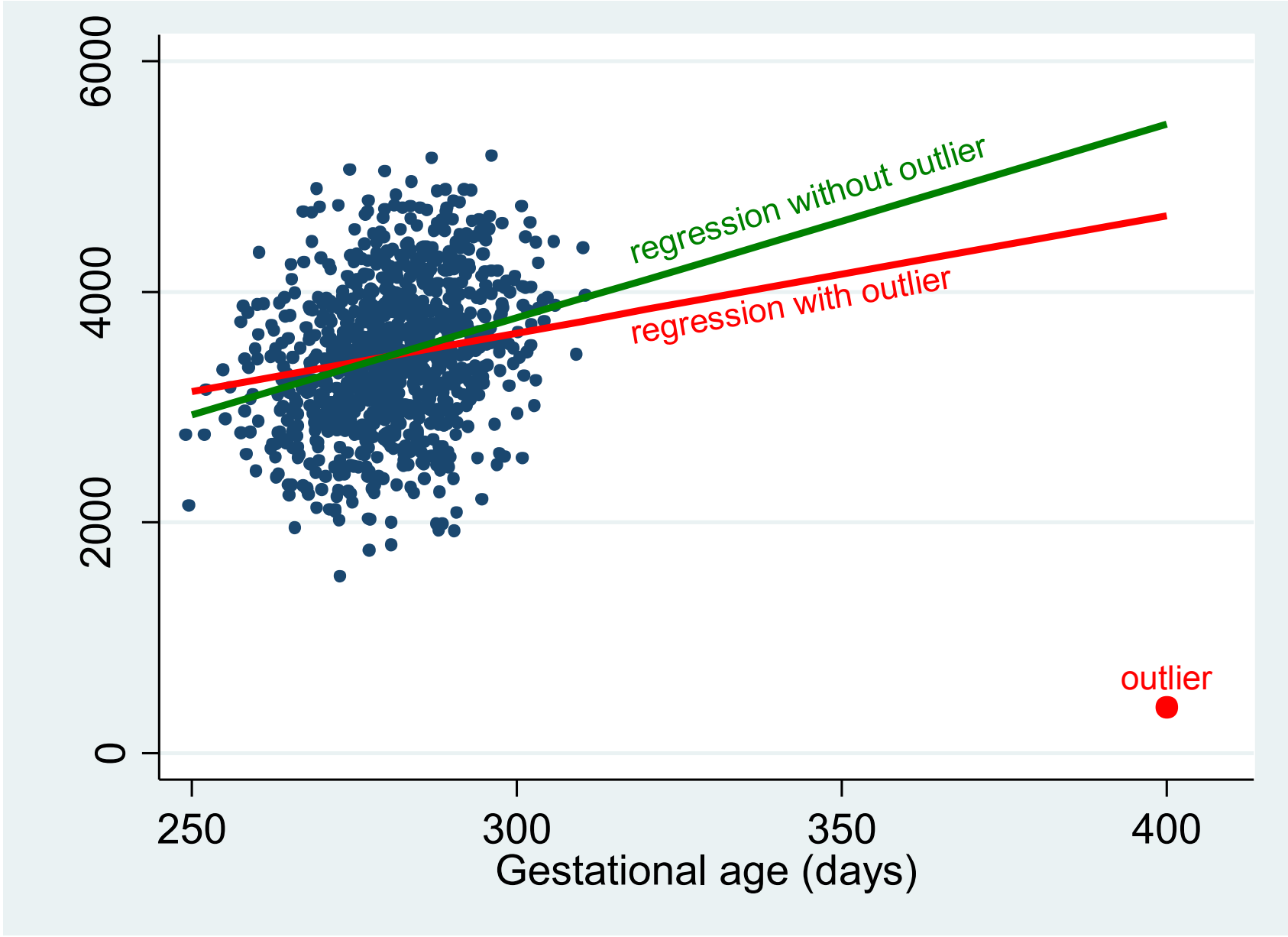
folkehelseinstituttet

Measures of influence

ROBUSTNESS

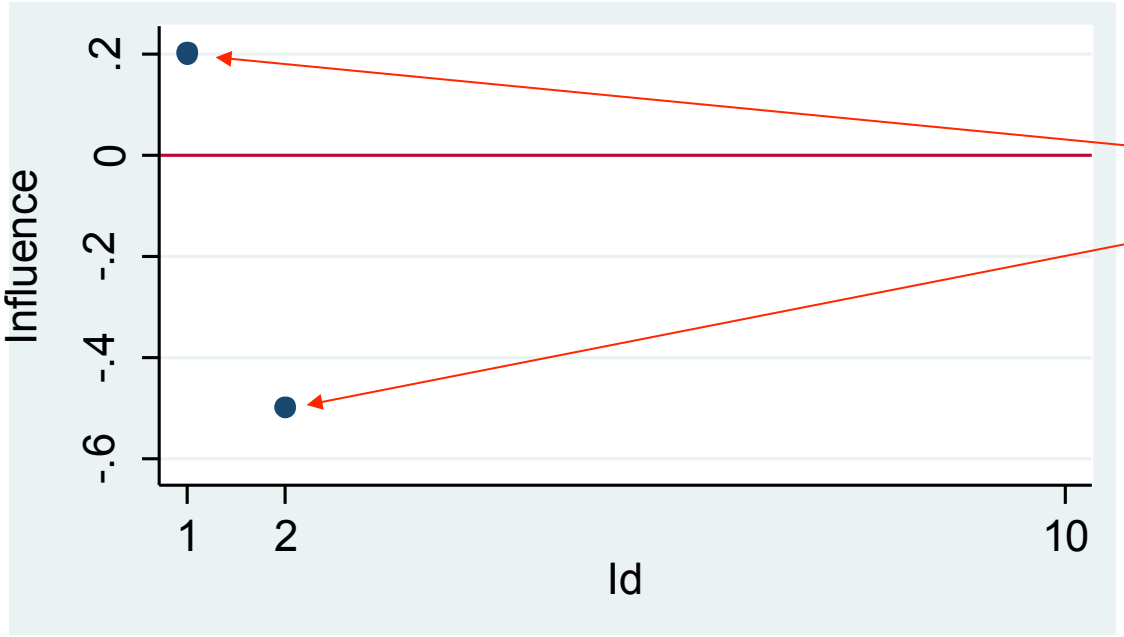
folkehelseinstituttet

Influence idea



folkehelseinstituttet

Measures of influence



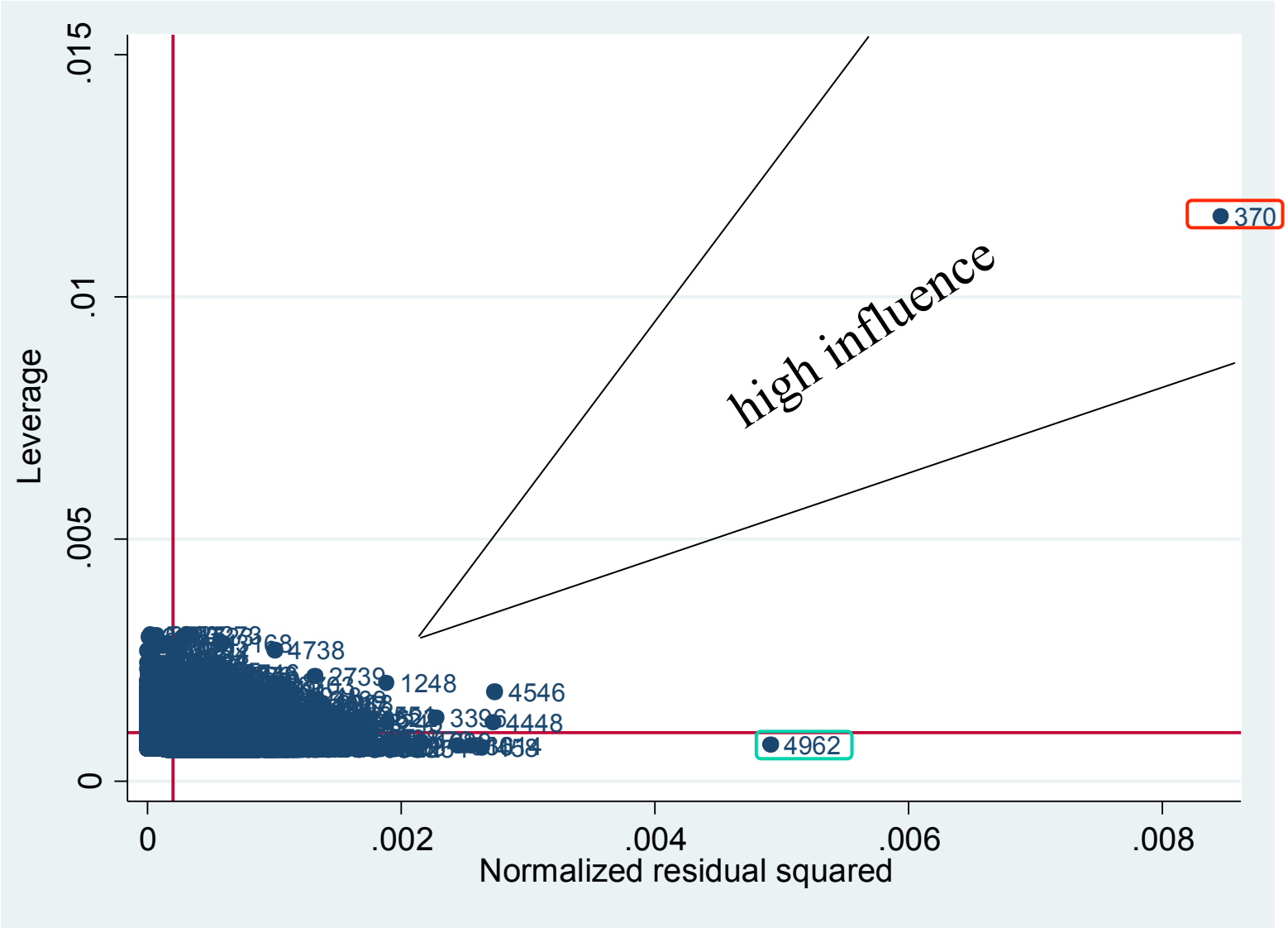
Remove obs 1, see change
remove obs 2, see change

- Measure change in:
  - Predicted outcome
  - Deviance
  - Coefficients (beta)
    - Delta beta

folkehelseinstituttet

Leverage versus residuals[2]

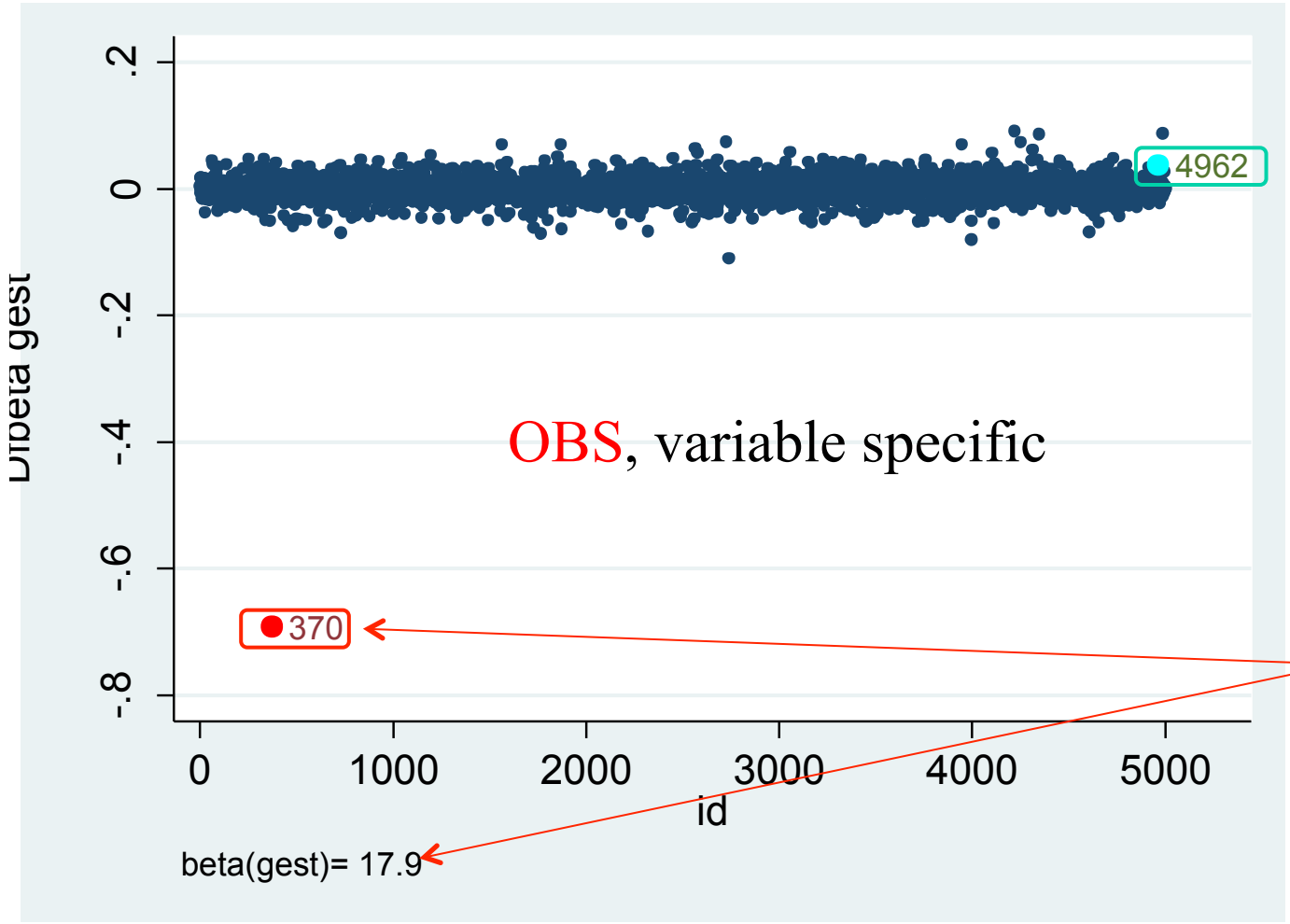
lvr2plot, mlabel(id)



folkehelseinstituttet

Delta-beta for gestational age

dfbeta(gest)
scatter _dfbeta_1 id



If obs nr 370 is removed, beta will change from 17.9 to 18.6

folkehelseinstituttet

Removing outlier

```
regress bw gest i.educ sex if id!=370
est store m4
est table m3 m4, b(%8.1f)
```

Variable	m3	m4
gest	17.9	18.5
educ		
2	71.5	64.2
3	99.1	88.6
sex	-154.3	-152.7
_cons	-1572.3	-1744.3

folkehelseinstituttet

Removing outlier

Full model N=5000

	coeff	95% conf. Int.
Birth weight at ref	3426	(3385 , 3467)
Gestational age		
per day	17.9	(16 , 20)
Education		
Low	0	
Medium	71.5	(25 , 118)
High	99.1	(51 , 148)
Sex		
Boy	0	
Girl	-154.3	(-187 , -121)

One outlier affected several estimates

Outlier removed N=4999

	coeff	95% conf. Int.
Birth weight at ref	3433	(3391 , 3474)
Gestational age		
per day	18.5	(17 , 20)
Education		
Low	0	
Medium	64.2	(18 , 110)
High	88.6	(40 , 137)
Sex		
Boy	0	
Girl	-152.7	(-185 , -120)

Final model

folkehelseinstituttet

Help

- Linear regression
  - help regress
    - syntax and options
  - help regress postestimation
    - dfbeta
    - estat hettest
    - lvr2plot
    - predict
    - margins

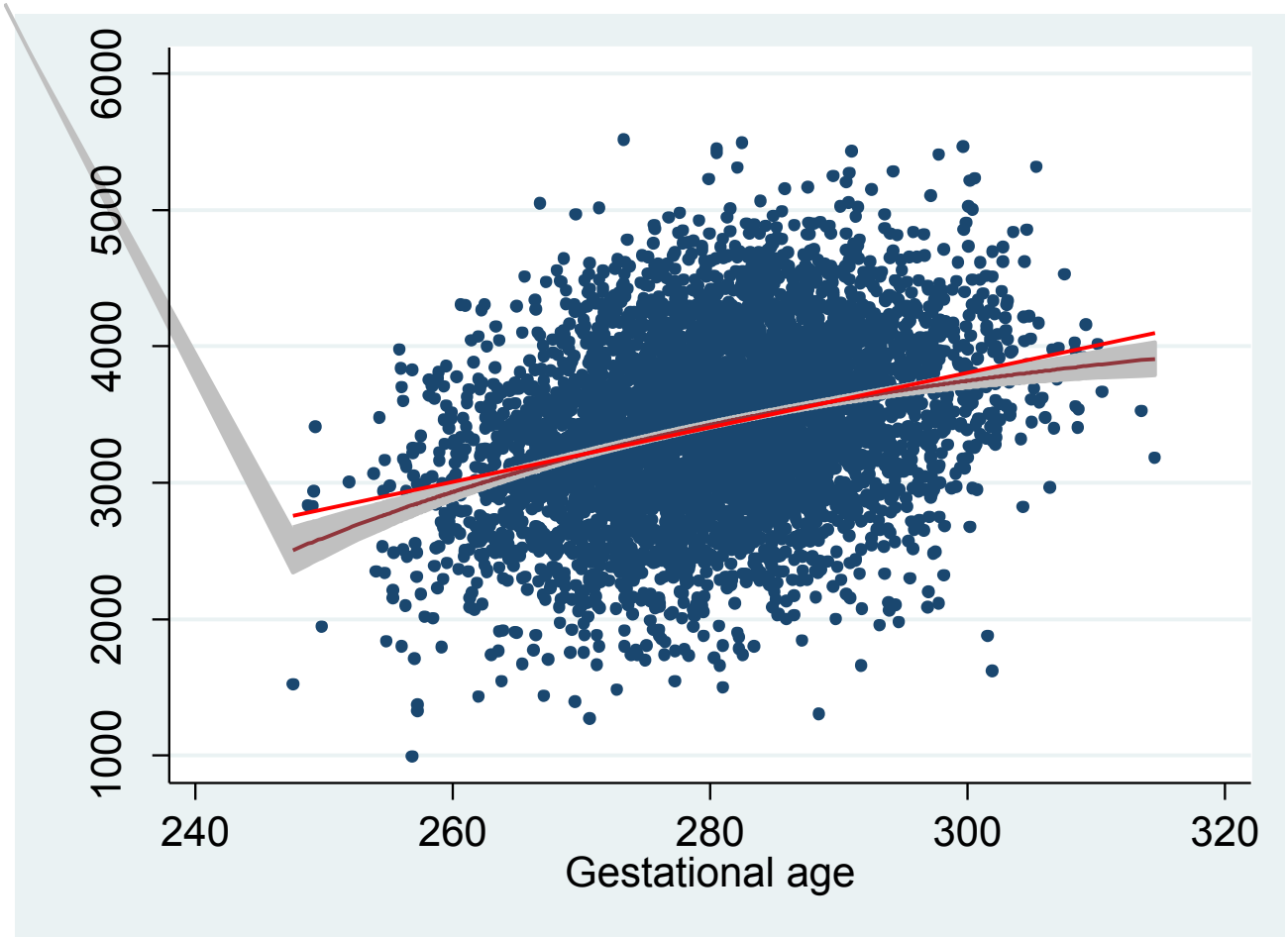
folkehelseinstituttet

bw2

NON-LINEAR EFFECTS

folkehelseinstituttet

bw2: Non-linear effects



Handle:
add
polynomial
or
spline

folkehelseinstituttet

Non-linear effects: polynomial

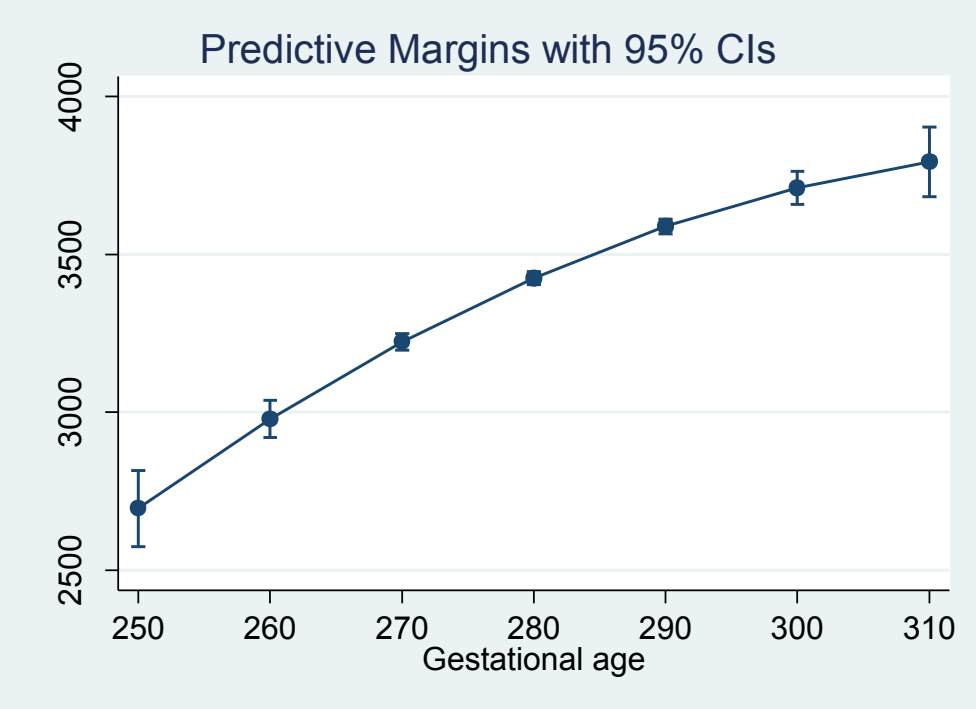
regress bw2 c.gest##c.gest i.educ sex

2. order polynomial in gest

bw2	Coef.	Std. Err.	t	P>|t|
gest	131.301	35.84957	3.66	0.000
c.gest#c.gest	-.2017914	.0638382	-3.16	0.002

margins, at(gest=(250(10)310))
marginsplot

predicted bw2 by gest



folkehelseinstituttet

Non-linear effects: spline

- Qubic spline

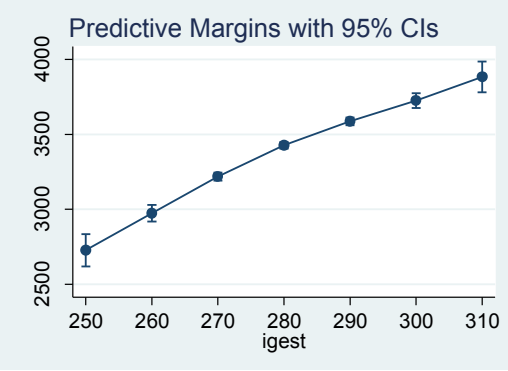
```
mkspline g=gest, cubic nknots(4)
regress bw2 g1 g2 g3 i.educ sex
```

make spline with 4 knots
regression with spline

- Plot

```
gen igest=5*round(gest/5)
margins, over(igest)
marginsplot
```

5-year integer values of gest
predicted bw



- Linear spline

```
mkspline g1 280 g2=gest
regress bw2 g1 g2 i.educ sex
```

make linear spline with knot at 280
regression with spline

bw2	Coef.	Std. Err.	t	P>|t|
g1	22.55257	1.801577	12.52	0.000
g2	14.01028	1.653815	8.47	0.000

folkehelseinstituttet

bw3

INTERACTION

folkehelseinstituttet

Interaction definitions

- Interaction: combined effect of two variables
- Scale
  - Linear models — additive
    - $y=b_0+b_1x_1+b_2x_2$ — both x_1 and x_2 = b_1+b_2
  - Logistic, Poisson, Cox — multiplicative
- Interaction
  - deviation from additivity (multiplicativity)

  - effect of x_1 depends on x_2

folkehelseinstituttet

bw3: Interaction (only linear effects)

- Add interaction terms

regress bw3 c.gest##i.sex i.educ — gest-sex interaction

bw3	Coef.	Std. Err.	t	P>|t|
gest	22.69672	1.276538	17.78	0.000
1.sex	2614.988	492.0968	5.31	0.000
sex#c.gest				
1	-9.89665	1.751459	-5.65	0.000

- Show results

margins, dydx(gest) at(sex=0) — effect of gest for boys
margins, dydx(gest) at(sex=1) — effect of gest for girls

	Delta-method dy/dx	Std. Err.	z	P>|z|
gest	22.69672	1.276538	17.78	0.000
gest	12.80007	1.308844	9.78	0.000

folkehelseinstituttet

Summing up 1

- Build model
  - regress bw gest — crude model
  - est store m1 — store
  - regress bw gest i.educ sex — full model
  - est store m2
  - est table m1 m2 — compare coefficients
- Interaction
  - regress bw3 c.gest##i.sex i.educ — test interaction
  - margins, dydx(gest) at(sex=0) — gest for boys
- Assumptions
  - predict res, residuals — residuals
  - predict pred, xb — predicted
  - scatter res pred — plot

folkehelseinstituttet

Summing up 2

- Non-linearity (linear spline)
  - mkspline g1 280 g2=gest — spline with knot at 280
  - regress bw2 g1 g2 i.educ sex — regression with spline
- Robustness
  - dfbeta(gest) — delta-beta
  - scatter _dfbeta_1 id — plot versus id