

Introduction to OLS with STATA

Erasmus+ visiting lecture at Faculty of
Management & Economics
University of Mons

Athanassios Stavrakoudis

<http://stavrakoudis.econ.uoi.gr>
Department of Economics, University of Ioannina, Greece

22 March 2017

Contents

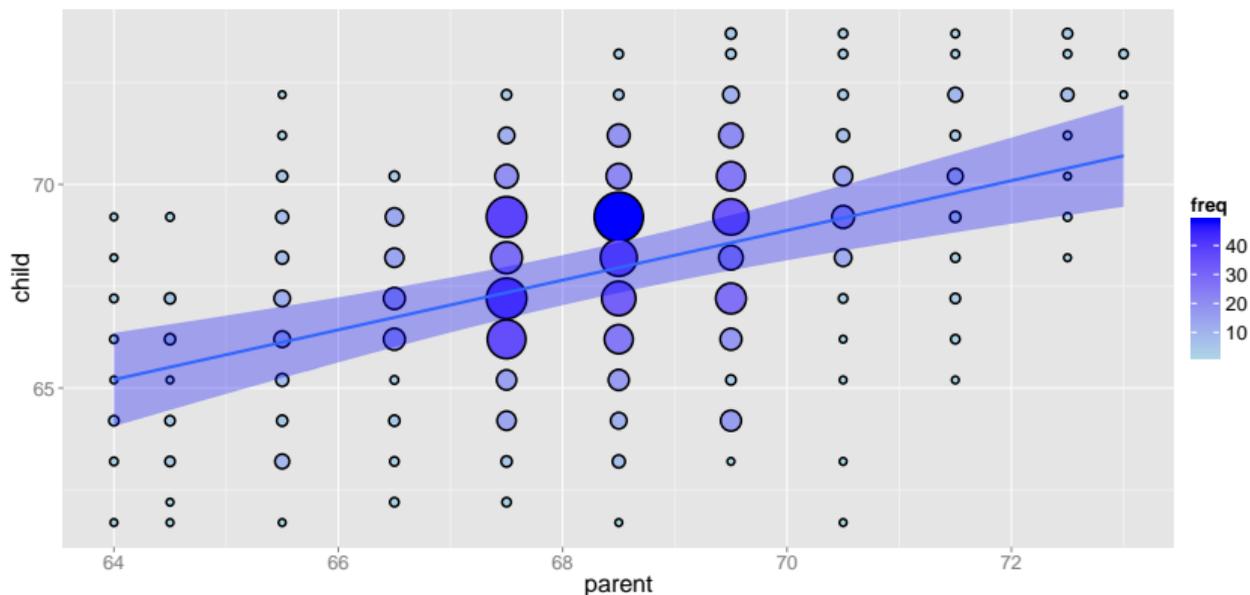
- 1 About this lecture
- 2 Ordinary Least Squares
- 3 Anscombe's quartet
- 4 Galton Example regression
- 5 Coefficient of determination
- 6 Confidence and Prediction Intervals

What is this about?

- An introduction basic linear models and Ordinary Least Squares with **STATA**.
- **STATA** is a software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.
- Download from : www.stata.com
- Various versions exist, **small STATA** is typically used for demo and classroom applications.
- There are plenty of books, internet resources and courses about **STATA**.
- **STATA** has gained big popularity in academia. especially to applied econometrics.

Some history

Over a century ago, F. Galton observed that children's height correlates with their parents height. This is one of the first application of OLS and linear models to science.



F. Galton, Regression towards mediocrity in hereditary stature, *The Journal of the Anthropological Institute of Great Britain and Ireland* **1886**, 15:246–263. doi:10.2307/2841583.

Variables in OLS method

 $Y =$ $X\beta + \epsilon$

Response variable

Response variable (or dependent variable) must be **continuous**.

Predictor variable

Predictor variable(s) (or independent variable) can be:

- 1 continuous
- 2 discrete
- 3 categorical

Data in OLS

Errors check/insect the data possible errors

Missing values Maybe some data values are missed

Patterns not all data sets are suitable for OLS

Outliers unusual or extreme values

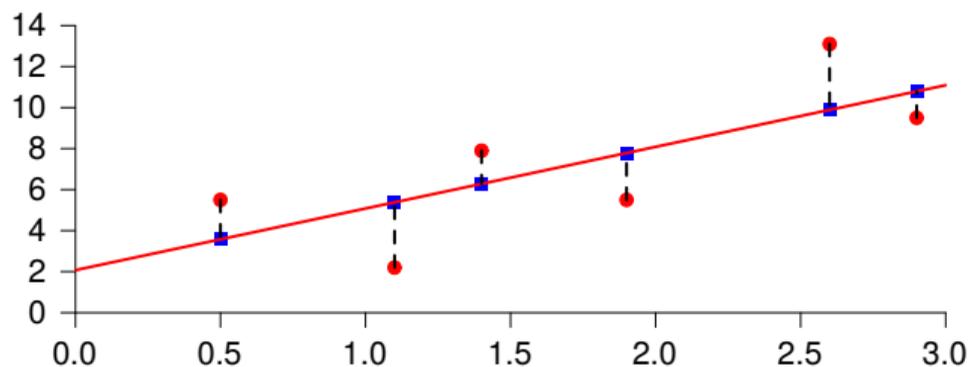
Plotting

Begin your analysis with plotting of data!

Contents

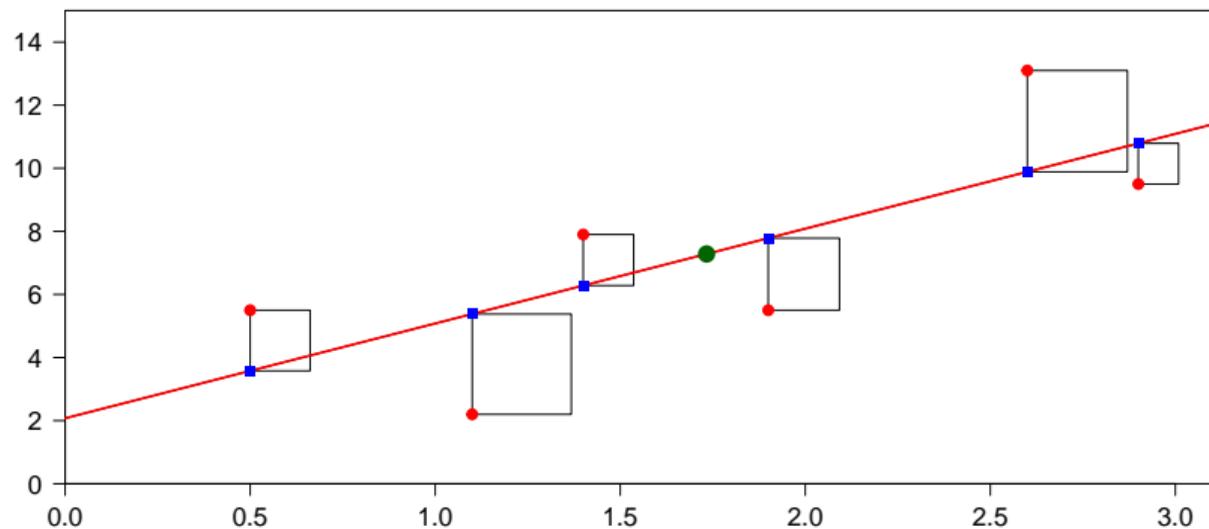
- 1 About this lecture
- 2 Ordinary Least Squares**
- 3 Anscombe's quartet
- 4 Galton Example regression
- 5 Coefficient of determination
- 6 Confidence and Prediction Intervals

Basic Terminology



- ① **red points:** $\{x_i, y_i\}$, observable variables (price, GDP, temperature, years in work, etc), out of line.
- ② **blue points:** $\{x_i, \hat{y}_i\}$, fitted (estimated) data, on the line..
- ③ **red line:** regression line (fitted from data)
- ④ **dashed lines:** residuals, $\{\hat{y}_i - y_i\}$ (estimation of errors)
- ⑤ **y:** outcome or dependent variable
- ⑥ **x:** explanatory variable or independent variable or regressor

Squared residuals



① **squares**: represent $(\hat{y}_i - y_i)^2$

② **Regression line** is the one that minimizes: $\sum_{i=1}^n (\hat{y}_i - y_i)^2$

③ **green point**: (\bar{x}, \bar{y}) , regression line passes through the mean values.

OLS, some reminders

$$\widehat{Y}_i = \beta_0 + \beta_1 \widehat{X}_i + \epsilon_i$$

$$E(\epsilon_i) = 0 \quad \forall i \quad \text{mean value of residuals}$$

$$V(\epsilon_i) = \sigma^2 \quad \forall i \quad \text{homoscedasticity}$$

$$\text{Cov}(\epsilon_i, \epsilon_j) = 0 \quad \forall i \neq j \quad \text{non correlated residuals}$$

$$\widehat{\beta}_1 = \text{cor}(Y, x) \frac{\sigma_Y}{\sigma_x} \quad \text{slope of the regression line}$$

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x} \quad \text{constant term}$$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} \quad \text{if } \beta_0 = 0$$

OLS, estimation in STATA cheatsheet

insheet using

read spreadsheet data (also .csv) from local or remote sites

summarize

Give descriptive statistics (mean, st.dev., etc)

generate

generate new variables using algebraic expressions

scatter

Scatter plot of two variables

histogram

Creates a histogram plot

graph

Create and export graphs and plots

regress y x

linear regression of y (response or dependent variable) on x (independent variable)

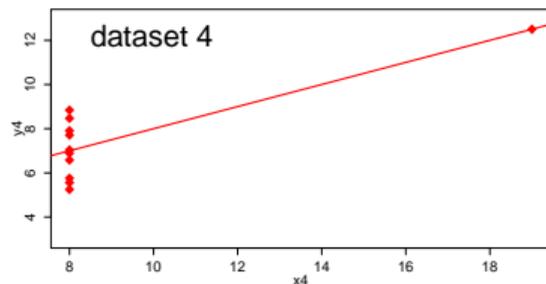
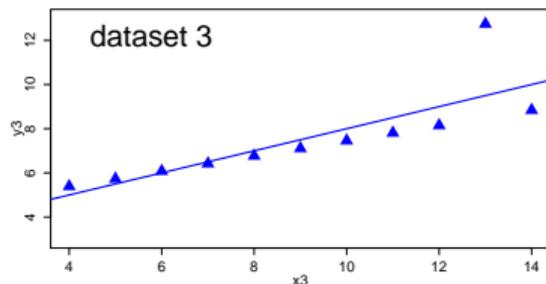
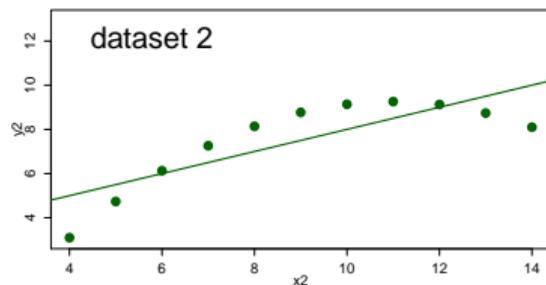
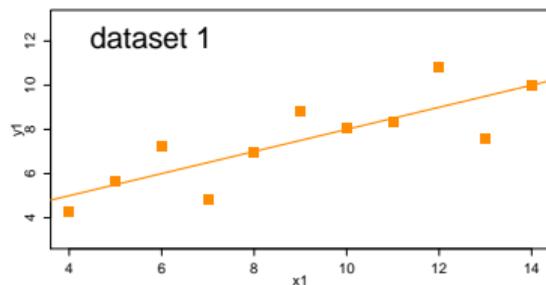
predict

create new variables like estimated \hat{y} and residuals \hat{e}

Contents

- 1 About this lecture
- 2 Ordinary Least Squares
- 3 Anscombe's quartet**
- 4 Galton Example regression
- 5 Coefficient of determination
- 6 Confidence and Prediction Intervals

Anscombe's quartet: four different data sets



F.J. Anscombe, Graphs in Statistical Analysis, *The American Statistician*, 1973, 27:17–21

```
1 clear
2 use http://www.stata-press.com/data/kk/anscombe
```

Same properties: mean and variance

```
1 summarize x1 x2 x3 x4
```

```
2
3   Variable |           Obs           Mean   Std. Dev.   Min   Max
4 -----+-----
5         x1 |             11              9   3.316625     4    14
6         x2 |             11              9   3.316625     4    14
7         x3 |             11              9   3.316625     4    14
8         x4 |             11              9   3.316625     8    19
```

```
9
10 summarize y1 y2 y3 y4
```

```
11
12  Variable |           Obs           Mean   Std. Dev.   Min   Max
13 -----+-----
14         y1 |             11   7.500909   2.031568     4.26  10.84
15         y2 |             11   7.500909   2.031657     3.1   9.26
16         y3 |             11              7.5   2.030424     5.39  12.74
17         y4 |             11   7.500909   2.030579     5.25  12.5
```

Attention!

All four data sets have the same mean and variance (to 2nd decimal point). This does not mean that they are samples of the same distribution!

Same properties: correlation and OLS coefficients

1	y1	Coef.	Std. Err.	t	P> t	Beta
2	-----+-----					
3	x1	.5000909	.1179055	4.24	0.002	.8164205
4	_cons	3.000091	1.124747	2.67	0.026	.
5	-----					

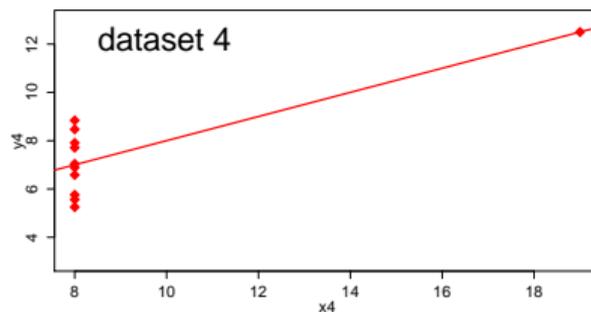
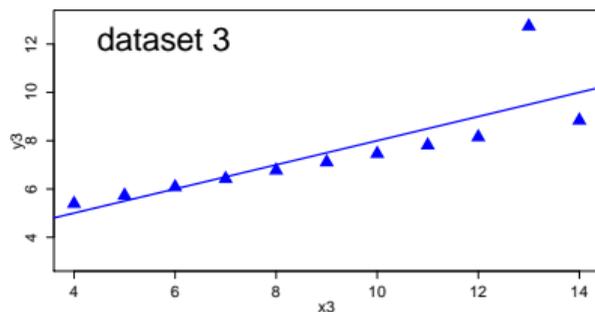
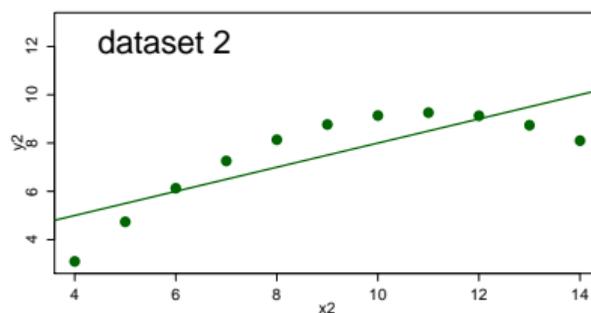
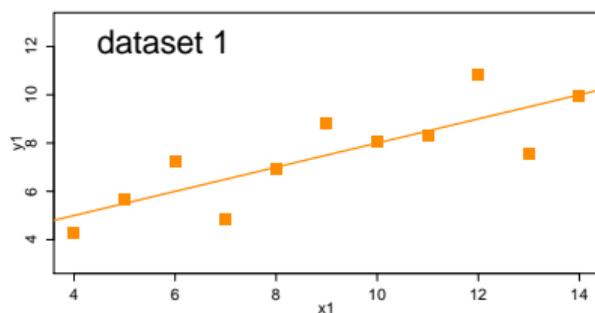
Attention!

Correlation and OLS coefficients are the same in all four data sets. This does not mean that imply that in all cases we !

Ancombe's quartet: plot your data

Do not use software as a black box

The four data sets have the same properties but obviously they are different: always plot your data as a first step of your analysis.



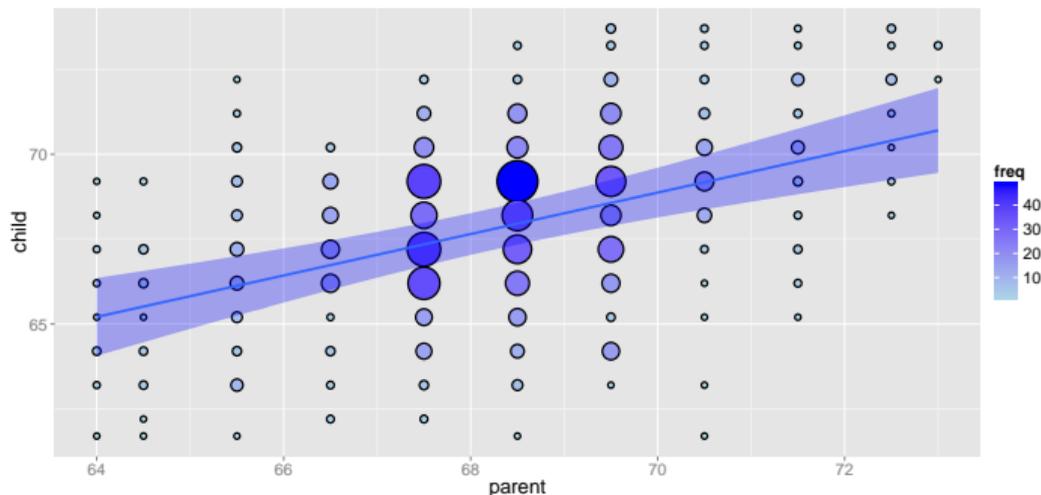
Question

In which case can we reasonably accept the linear model hypothesis?

In which case can we apply OLS method to analyze the data?

Galton regression

$$\text{childHeight} = \beta_0 + \beta_1 \text{parentHeight} + \epsilon$$



Galton, F. "Regression towards mediocrity in hereditary stature". *The Journal of the Anthropological Institute of Great Britain and Ireland* **1886**, 15:246–263. doi:10.2307/2841583.

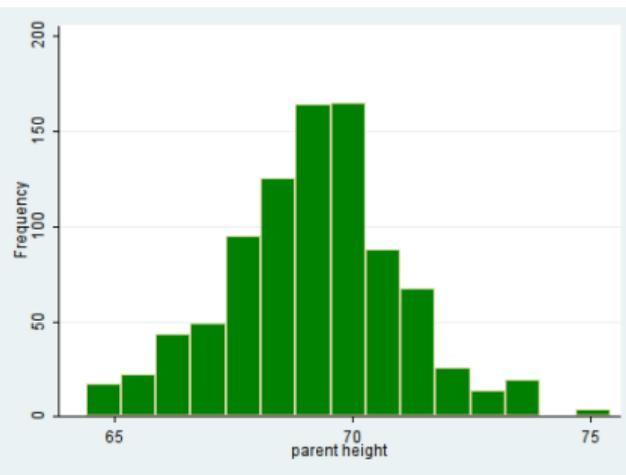
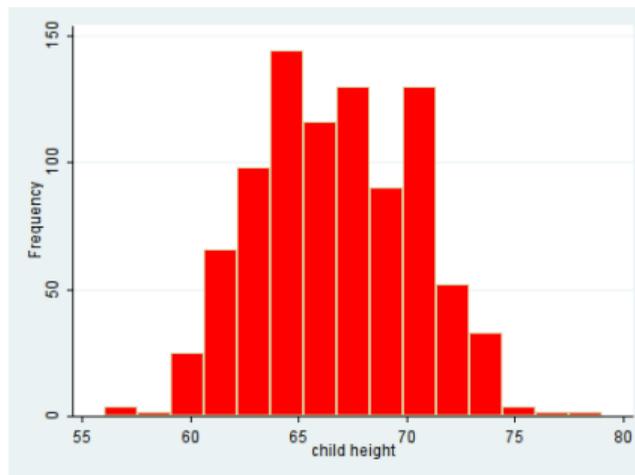
Galton Data

```
1 insheet using http://www.math.uah.edu/stat/data/Galton.csv
2 summarize father mother height
3 generate parent = 0.5 * (father + 1.08*mother)
4 generate child = height
5 summarize parent child
```

$$parent = \frac{1}{2}(father + 1.08 \times mother)$$

- 1 Galton's data is one of the most famous dataset in the world.
- 2 We create a new variable (*parent*) to represent parent's height (both father and mother, pay attention to the weighted mean).
- 3 The column height in the original data set represent the child's height.

Histograms

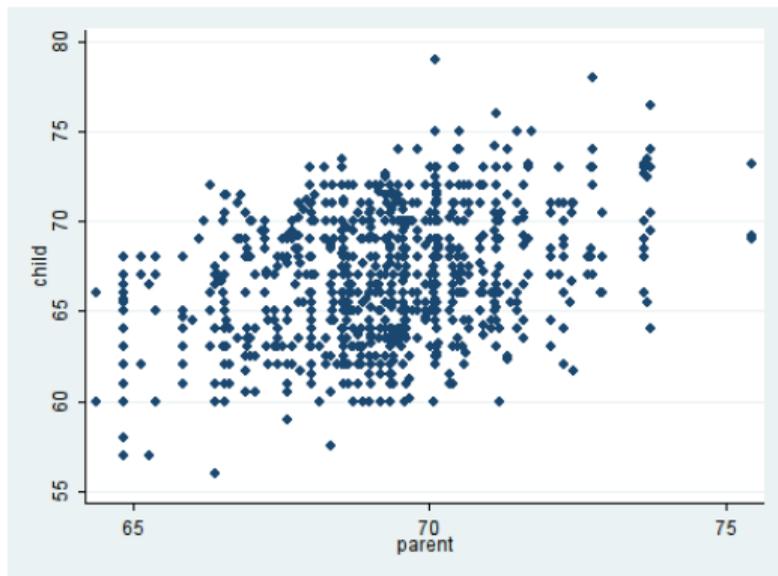


```

1 histogram child, bin(15) frequency fcolor(red) xtitle("child height")
2 graph export child-hist.png
3 histogram parent, bin(15) frequency fcolor(green) xtitle("parent height")
4 graph export parent-hist.png

```

Scatter plot



Scatter plot

What is the relationship between the heights parents and the heights of children?

A scatter plot always helps to have a visual perception of the correlation between two variables.

- 1 `graph twoway scatter child parent`
- 2 `graph export scatter1.png, replace`

Child's height vs parent's height, STATA output

```

1 regress child parent
2
3           Source |           SS           df           MS           Number of obs =           898
4 -----+-----
5           Model |    1218.53556           1    1218.53556           F( 1, 896) =    106.04
6           Residual |   10296.5259           896    11.4916584           Prob > F           =    0.0000
7 -----+-----
8           Total |   11515.0615           897    12.8373038           R-squared           =    0.1058
9                                     Adj R-squared       =    0.1048
10                                     Root MSE           =    3.3899
11
12 -----+-----
13           child |           Coef.   Std. Err.       t       P>|t|       [95% Conf. Interval]
14 -----+-----
15           parent |    .6411904    .0622672     10.30    0.000       .5189839       .7633969
16           _cons |    22.3762    4.311744      5.19    0.000       13.91391       30.8385

```

Child's height vs parent's height

Review

```

4 summarize father mother height
5 generate parent = 0.5 * (father +
6 generate child = height
7 summarize parent child
8 histogram child, bin(15) frequency
9 graph export child-hist.png, replac
10 histogram parent, bin(15) frequen
11 graph export parent-hist.png, repl
12 graph twoway scatter child parent
13 graph export scatter1.png, replac
14 correlate child parent
15 regress child parent

```

Variables

Name	Label	Type	Format
family	Family	str4	%3s
father	Father	float	%3.0g
mother	Mother	float	%3.0g
gender	Gender	str1	%3s
height	Height	float	%3.0g
kids	Kids	byte	%8.0g
parent		float	%3.0g
child		float	%3.0g

Command

```

. regress child parent

```

Source	SS	df	MS
Model	1218.53556	1	1218.53556
Residual	10296.5259	896	11.4916584
Total	11515.0615	897	12.8373038

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
child						
parent	.6411904	.0622672	10.30	0.000	.5189839	.7633969
_cons	22.3762	4.311744	5.19	0.000	13.91391	30.8385

Number of obs = 898
F(1, 896) = 106.04
Prob > F = 0.0000
R-squared = 0.1058
Adj R-squared = 0.1048
Root MSE = 3.3899

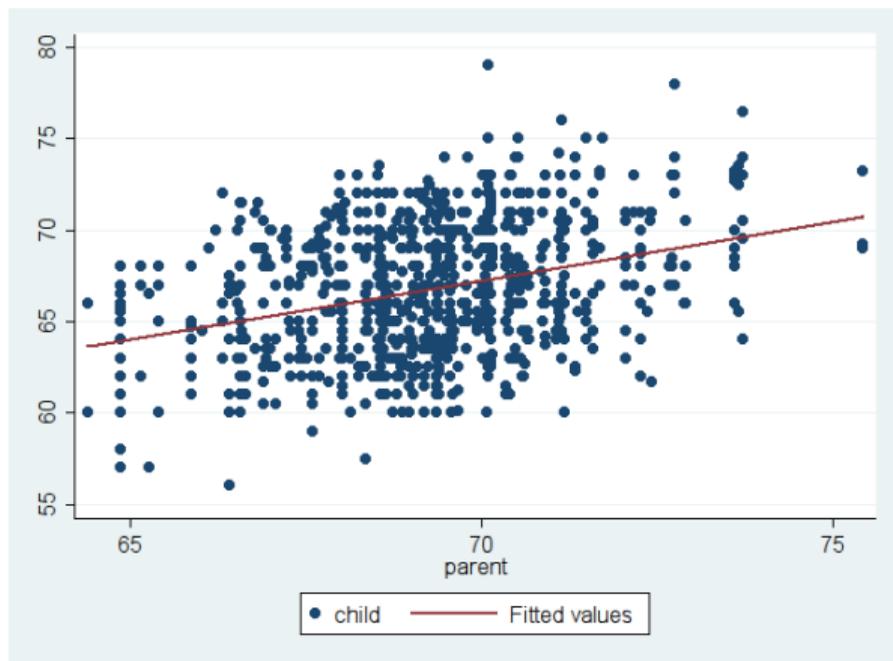
Z:\home\astavrak\edu\erasmus\lectures\Mons2017\OLS\stata\examples CAP NUM OVR

Interpretation

- 1 A child's height is $\approx 23.9''$ + 0.64 times parent height.
- 2 A child's height is $\approx 23.9''$ if his/her parent height is $\approx 0''$.
- 3 If two parents height differ by 1" then their children height will differ by $\approx 0.64''$

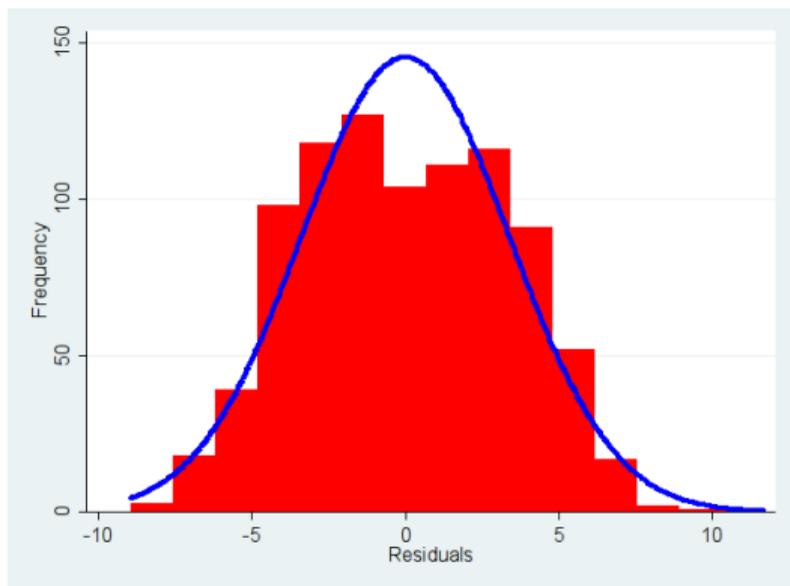
1 **regress** child parent

Scatter plot with regression line



```
1 graph twoway (scatter child parent) (lfit child parent)
2 graph export scatter2.png, replace
```

Residuals



Normality

We assume that residuals distribute normally

```
1 predict residuals, resid  
2 histogram residuals, frequency bin(15) color(red) normal
```

Residuals, skewness, kurtosis, normality

```
1 sktest residuals
```

```
2
```

```
3 Skewness/Kurtosis tests for Normality
```

```
4 ----- joint -----
```

Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	Prob>chi2
residuals	898	0.4555	0.0000	37.37	0.0000

```
8
```

```
9 . swilk residuals
```

```
10
```

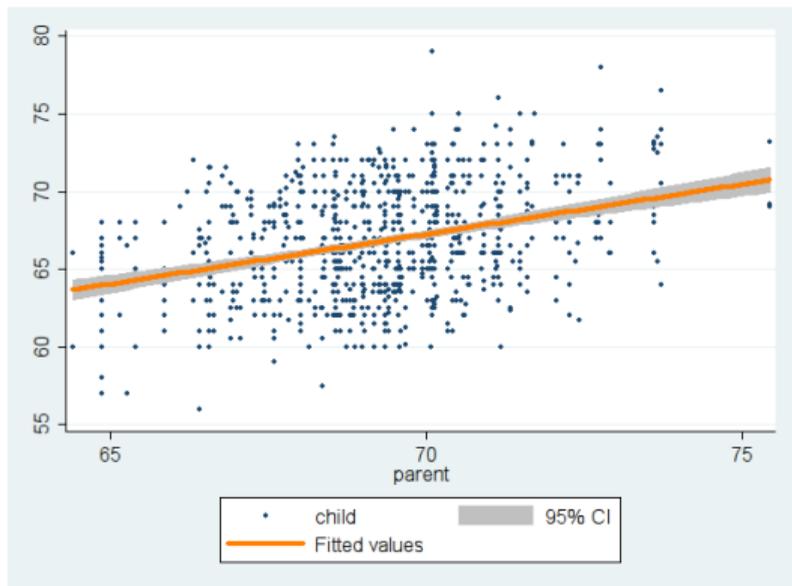
```
11 Shapiro-Wilk W test for normal data
```

```
12
```

Variable	Obs	W	V	z	Prob>z
residuals	898	0.98722	7.305	4.903	0.00000

```
15
```

confidence interval plot



Scatter plot

```
1 graph twoway (scatter child parent) (lfitci child parent)  
2 graph export scatter2ci.png, replace
```

Coefficients, standard errors and t-statistics

	(1)	(2)
	child	child
parent	0.641*** (10.30)	0.641*** (10.30)
_cons	22.38*** (5.19)	22.38*** (5.19)
<i>N</i>	898	898

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Standard error of the coefficient

$$s_{\hat{\beta}_1} = \sqrt{\frac{\frac{1}{n-2} \sum \hat{\varepsilon}_i^2}{\sum (x_i - \bar{x})^2}}$$

Standard error of the coefficient

General Model: $Y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$

Galton: $child_i = \beta_0 + \beta_1 \cdot parent_i + \epsilon_i$

$$H_0 : \beta_1 = 0 \quad \longleftrightarrow \quad H_1 : \beta_1 \neq 0$$

T-test statistic:

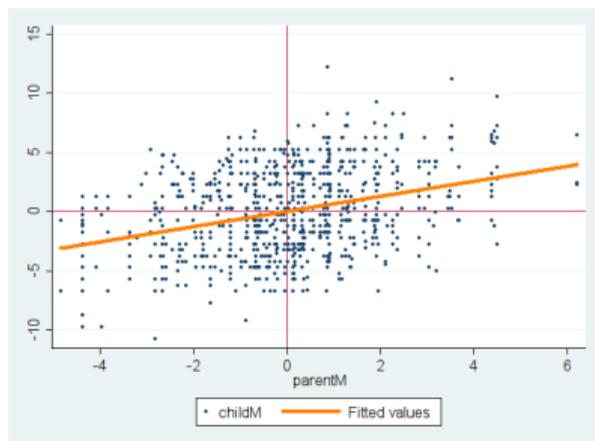
$$T_1 = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\beta_1}}$$

With normality of ϵ_i under the Null hypothesis:

$$T_1 \sim t_{n-2}$$

$$\text{p-value: } p = 2 \cdot (1 - F_{t_{n-2}}(|T_1|))$$

Regression to the mean



Subtract the mean:

$$childM = child - \overline{child}$$

$$parentM = parent - \overline{parent}$$

and estimate this:

$$childM = \beta_0 + \beta_1 parentM + \varepsilon$$

```

1 summarize child, meanonly
2 generate childM = child - r(mean)
3 summarize parent, meanonly
4 generate parentM = parent - r(mean)
5 summarize childM parentM
6 regress childM parentM
7 graph twoway (scatter childM parentM) (lfit childM parentM)
8 graph export scatter3.png, replace

```

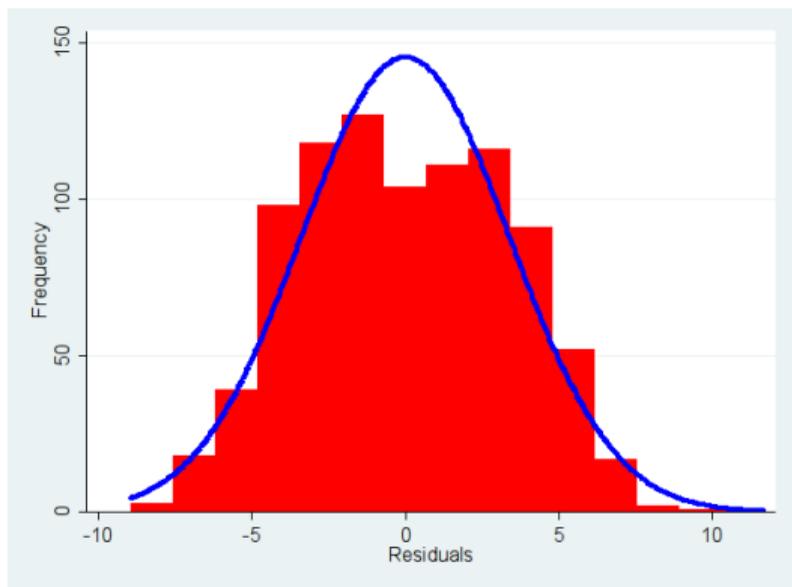
Regression to the mean, STATA output

```

1 . regress childM parentM
2
3           Source |           SS           df           MS           Number of obs =           898
4 -----+-----
5           Model |    1218.53556           1    1218.53556           F( 1, 896) =    106.04
6           Residual |   10296.5259           896    11.4916584           Prob > F           =    0.0000
7 -----+-----
8           Total |   11515.0615           897    12.8373037           R-squared           =    0.1058
                                           Adj R-squared       =    0.1048
                                           Root MSE           =    3.3899
9
10 -----+-----
11          childM |           Coef.   Std. Err.       t       P>|t|       [95% Conf. Interval]
12 -----+-----
13          parentM |    .6411904     .0622672     10.30    0.000     .5189839     .7633969
14          _cons |    1.13e-08     .1131236     0.00    1.000    -.2220181     .2220181
15 -----+-----

```

Residuals, regression to the mean

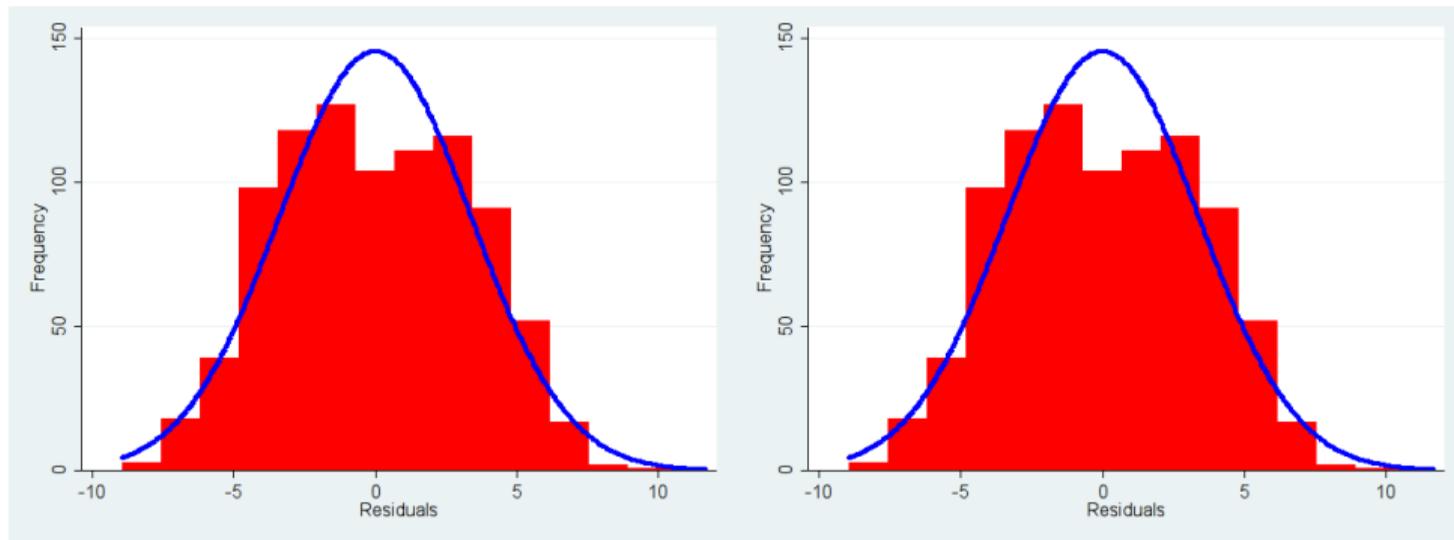


Normality

We assume that residuals distribute normally

```
1 predict residualsM, resid
2 histogram residualsM, frequency bin(15) color(red) normal
```

Identical plots? Why?



Is it normal?

Left: original regression, Right: Regression to the mean.

Can you spot any difference in these plots?

Is it normal that both regressions result to identical residual distribution?

Contents

- 1 About this lecture
- 2 Ordinary Least Squares
- 3 Anscombe's quartet
- 4 Galton Example regression
- 5 Coefficient of determination**
- 6 Confidence and Prediction Intervals

Coefficient of determination R^2

The coefficient of determination (**Multiple R-squared** in **R**) indicates the proportion of the variance in Y , which is explained by the model.

$$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{ESS}{TSS} = \frac{S_{xY}^2}{S_x^2 S_Y^2} = r_{xY}^2$$

$R^2 \rightarrow 1$ indicates a good fit

$R^2 \rightarrow 0$ indicates a poor fit

$$TSS = \sum (y_i - \bar{y})^2$$

Total sum of squares

$$ESS = \sum (y_i - \hat{y})^2$$

Error sum of squares

$$RSS = \sum (\hat{y}_i - \bar{y})^2$$

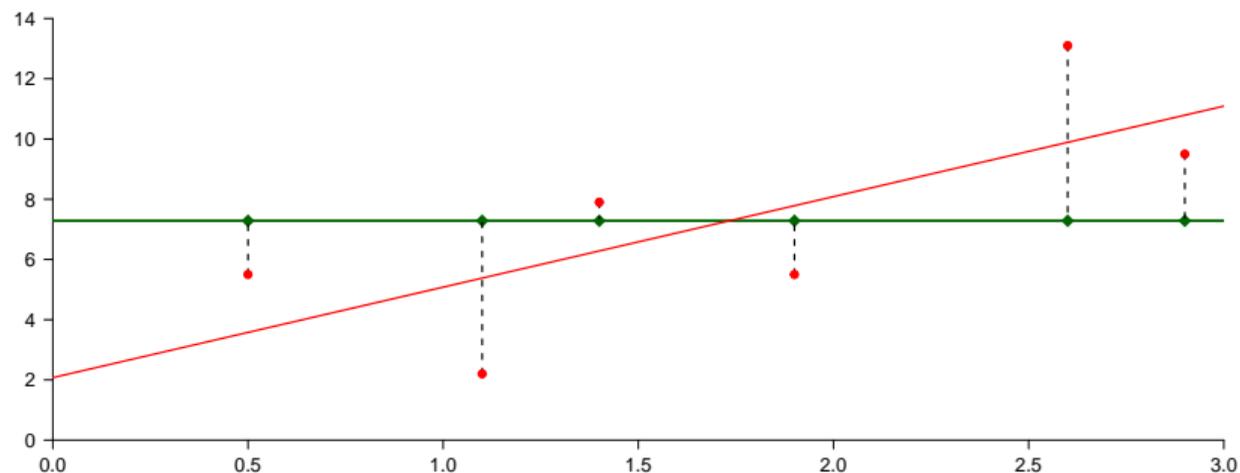
regression sum of squares

$$TSS = RSS + ESS$$

$$R^2 = RSS/TSS$$

Fraction of unexplained variance (FUV)

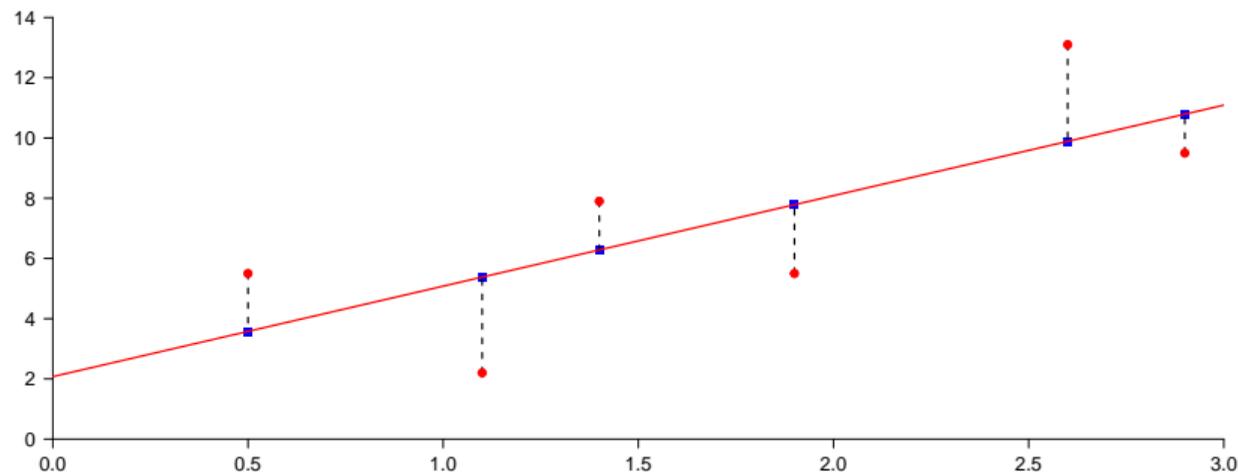
Total sum of squares (TSS)



$$TSS = \sum (y_i - \bar{y})^2$$

- ① **red points:** y_i , observations
- ② **green points:** \bar{y} , mean value
- ③ **dashed lines:** $y_i - \bar{y}$

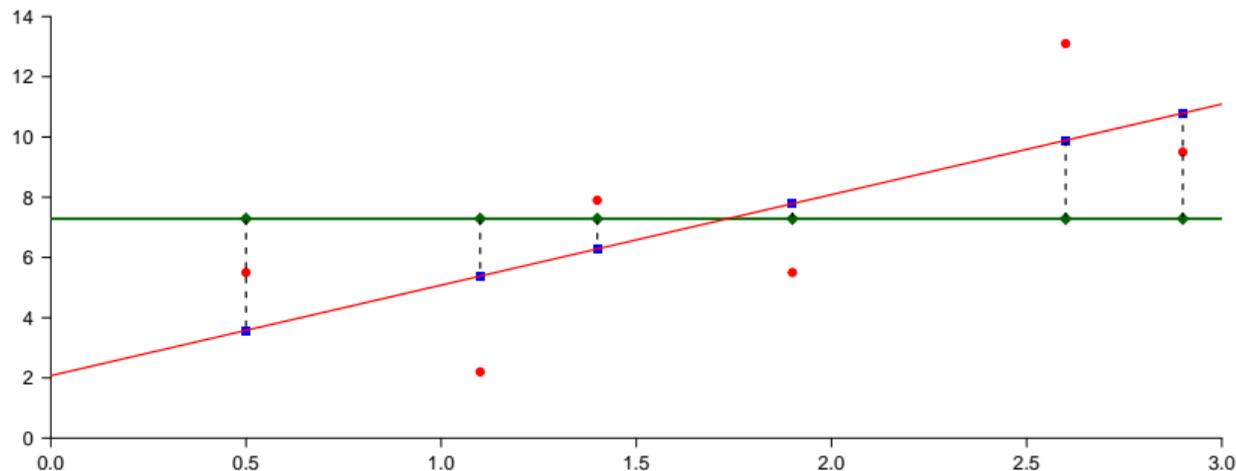
Error sum of squares (ESS)



$$\begin{aligned} ESS &= \sum (y_i - \hat{y}_i)^2 = \sum \epsilon_i^2 \\ &= \sum \left(y_i - (\beta_0 + \beta_1 x_i) \right)^2 \end{aligned}$$

- ① **red points:** y_i , observations
- ② **blue points:** \hat{y}_i , fitted values
- ③ **dashed lines:** $y_i - \hat{y}_i$

Regression sum of squares (RSS)



$$RSS = \sum (\hat{y} - \bar{y})^2$$

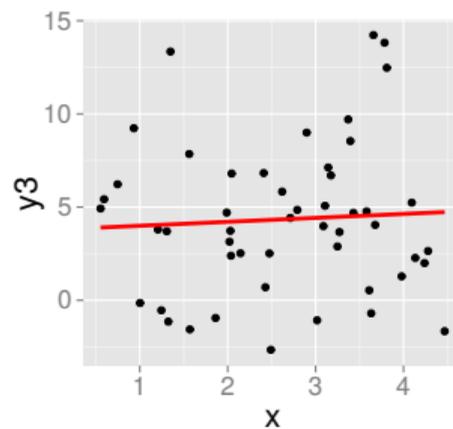
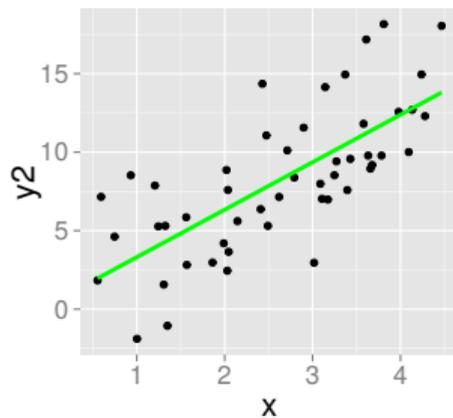
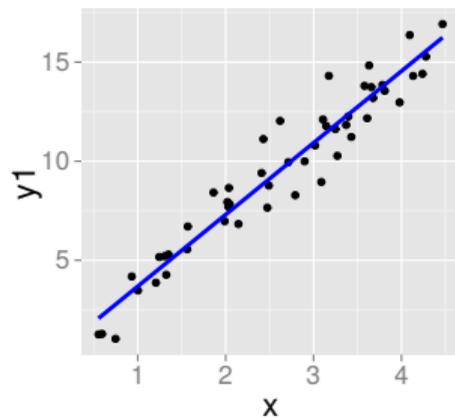
- ① **red points:** y_i , observations
- ② **blue points:** \hat{y}_i , fitted values
- ③ **green points:** \bar{y} , mean value of y
- ④ **dashed lines:** $\hat{y} - \bar{y}$

R^2 value

$$R^2 \equiv 1 - \frac{ESS}{TSS} = \frac{RSS}{TSS}$$

$$\bar{R}^2 = 1 - \frac{n-1}{n-k} (1 - R^2)$$

Comparison with R2



R^2 values

Bigger R^2 values **indicate** better fit of the data.

Caution: it is not a strong criterion.

Be careful of R^2 inflation.

Contents

- 1 About this lecture
- 2 Ordinary Least Squares
- 3 Anscombe's quartet
- 4 Galton Example regression
- 5 Coefficient of determination
- 6 Confidence and Prediction Intervals**

Predicting in linear models

Point prediction

Suppose we have a point $parent = 71.5$. We want to know what is the predicted value by the model. This is:

$$child = \hat{\beta}_0 + \hat{\beta}_1 \times parent$$

The following code makes the needed computation.

```

1 predcalc child, xvar(parent=71.5)
2
3 Model:      Linear Regression
4 Outcome:    -- child
5 X Values:   parent=71.5
6 Num. Obs:   898
7
8 Predicted Value and 95% CI for child:
9
10           68.22 ( 67.87, 68.58)

```

Point estimation results

Try to interpret the results.

What these 68.22 (67.87, 68.58) values tell us?

Prediction intervals STATA

```

1 regress child parent
2 predict Yhat
3 predict CIstderror, stdp
4 predict PIstderror, stdf
5 generate tmult = invttail(896, .025)
6 generate lowerCI = Yhat - tmult*CIstderror
7 generate upperCI = Yhat + tmult*CIstderror
8 generate lowerPI = Yhat - tmult*PIstderror
9 generate upperPI = Yhat + tmult*PIstderror
10 list Yhat lowerCI upperCI lowerPI upperPI in 1

```

11

```

12 +-----+
13 |      Yhat      lowerCI      upperCI      lowerPI      upperPI |
14 |-----|
15 1. |  70.7412   69.94581   71.53658   63.99964   77.48275 |
16 |-----|

```

Prediction interval

fit An lm object

newdata data for prediction (must be in data frame format)

interval can be **"confidence"** for confidence interval or **"prediction"** for prediction interval

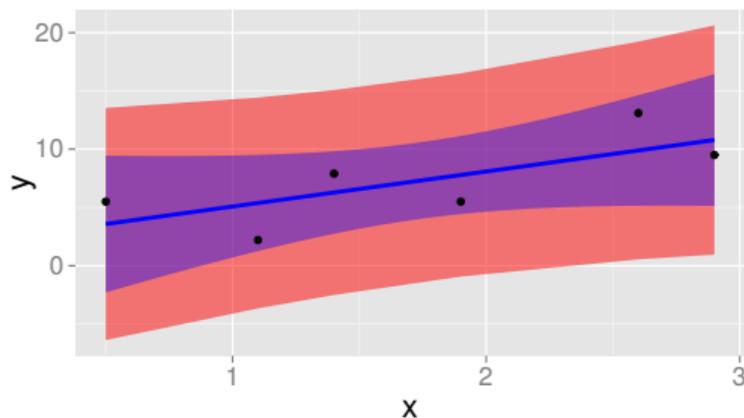
level tolerance/confidence level

Under the assumption of normally distributed residuals the prediction interval can be calculated as:

$$\hat{Y}_P = \hat{Y} \pm \hat{\sigma} \sqrt{1 + \frac{1}{n} \left(1 + \frac{(x_P - \bar{x})^2}{\tilde{S}_x^2} \right)} \cdot t_{n-2; 1-\alpha/2}$$

with $\hat{\sigma} = \frac{1}{n-2} \cdot \sum \hat{\epsilon}_i^2$ and $\tilde{S}^2 = \frac{n-1}{n} \cdot S^2$

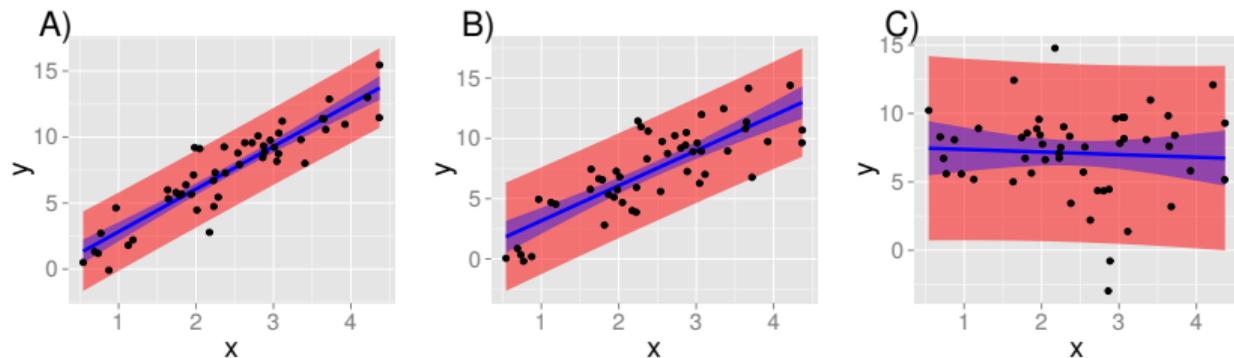
Confidence interval vs Prediction interval



Notes and explanations

- 1 Confidence interval tells us how good is the fitting.
- 2 Prediction interval tells where new data points are expected.
- 3 **Confidence** : about existed data. **Prediction** : about non existed data
- 4 Prediction interval area is always wider than confidence interval area.

Confidence interval vs Prediction interval



Discussion

Examine the three plots above and comment on the following:

- 1 Variance of y and ϵ
- 2 R^2
- 3 Std. Error of the coefficient (slope, $\widehat{\beta}_1$)
- 4 P-value of the coefficient (slope, $\widehat{\beta}_1$)
- 5 Band width of confidence interval

Comments and Questions

Thank you very much
for your attention!