Multiple regression

Statistical Reasoning
and Quantitative Methods

François Briatte & Ivaylo Petev

Session 10

Outline

Multiple regression modelling expands simple linear regression to any number and type of variables, and provides interpretable **parameters** in the form of regression coefficients.

Regression modelling

Regression coefficients

Our course stops here, as it gets more difficult afterwards.

We will explore diagnostics next week and then finish the course with logistic regression.

Fitting a **simple** linear regression model

The basic model fits a **linear function** to the data, of the form:

$$Y = \alpha + \beta X + \epsilon \text{ or identically } \hat{Y} = \alpha + \beta X$$

where:

- Y is the **dependent variable** (response)
- X is the **independent variable** (predictor)
- α is the **constant** (intercept)
- β is the **regression coefficient** (slope)
- ϵ is the **error term** (residuals)

Note: the model assumes that the relationship is **linear**.

Fitting a **multiple** linear regression model

The model also fits a **linear function** to the data, of the form:

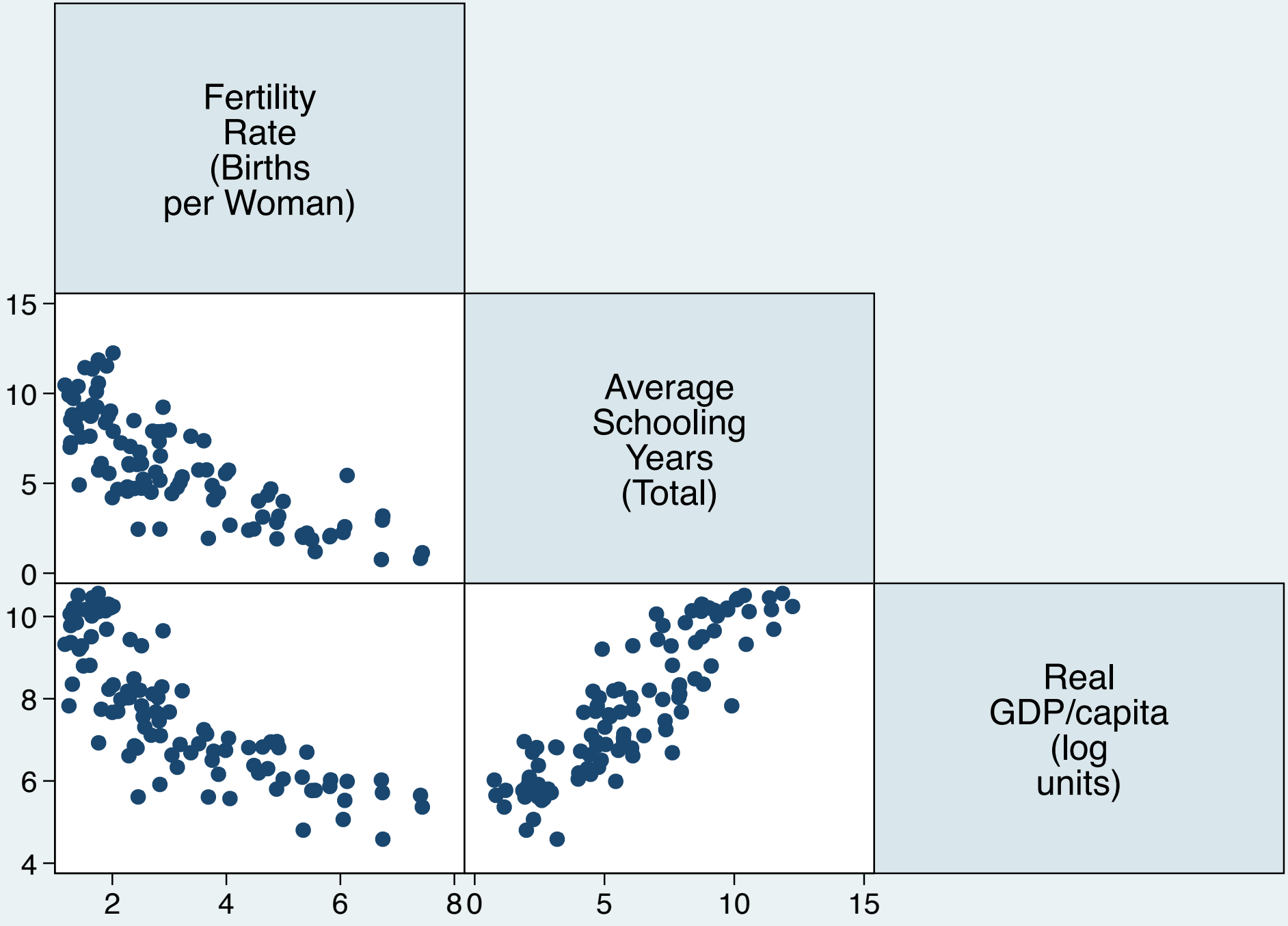
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

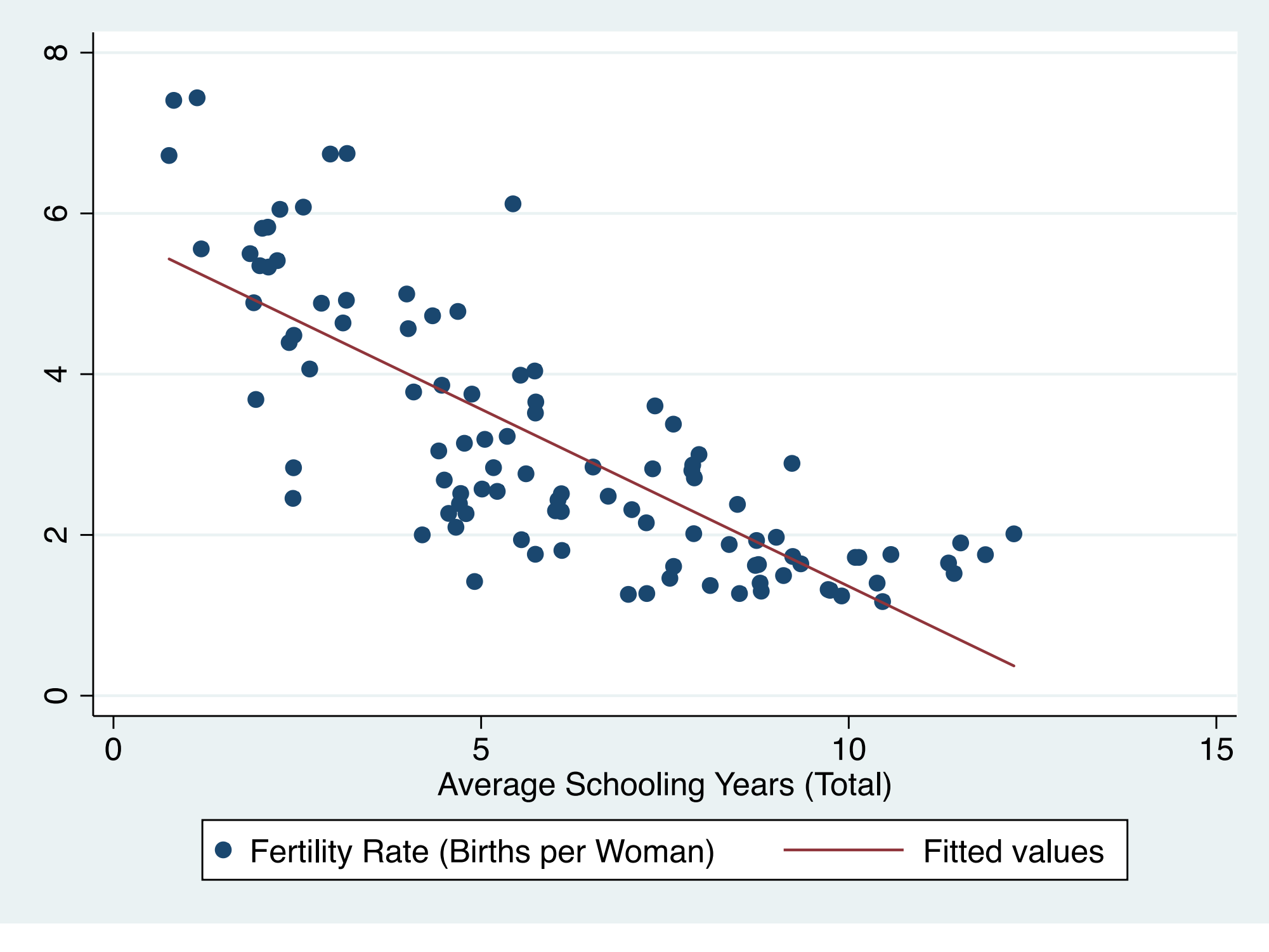
where:

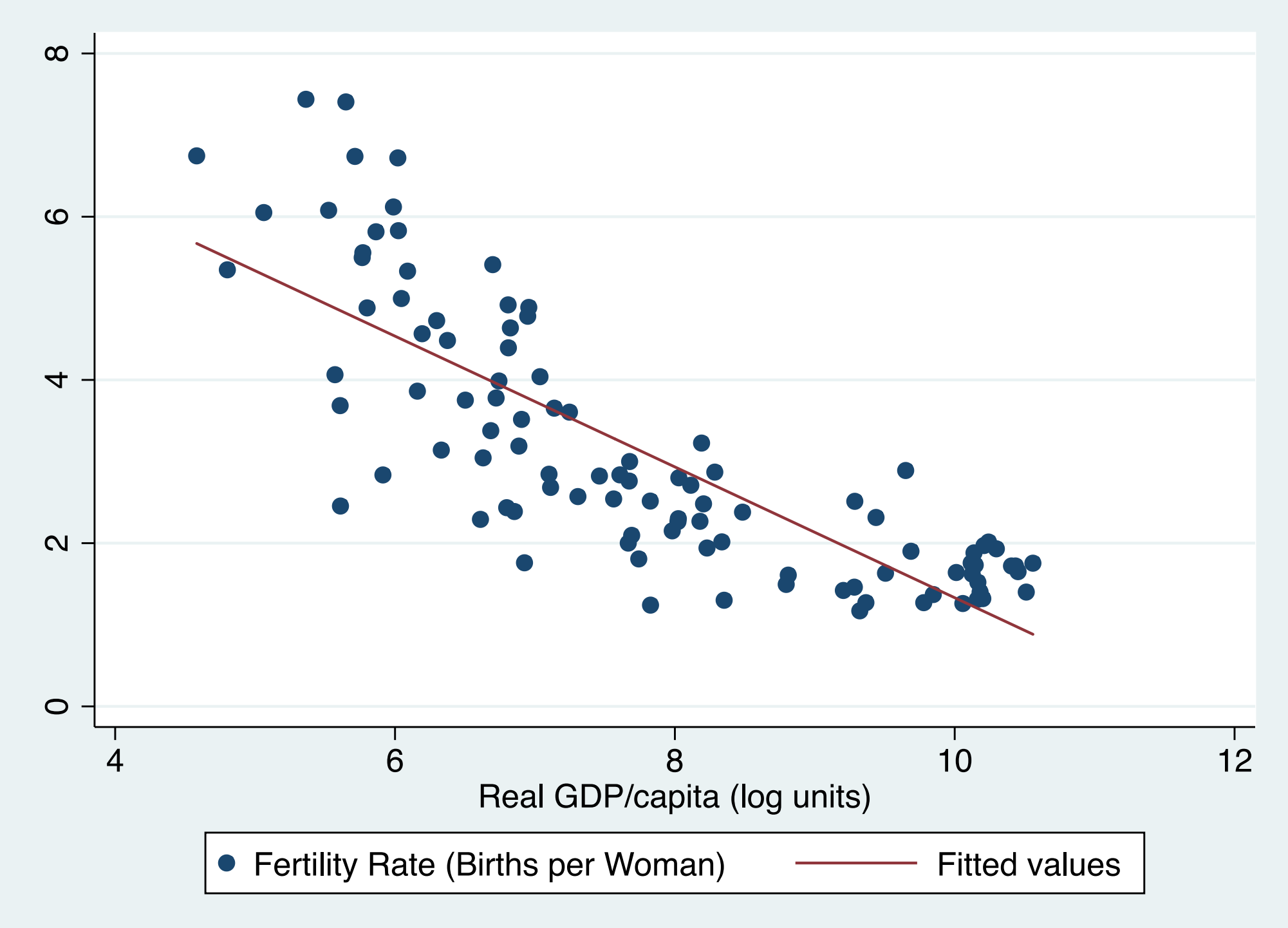
- Y is the **dependent variable** (response)
- X is a **vector** of **independent variables** (predictors)
- α is the **constant**
- $\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$ is a **vector** of **regression coefficients**
- ϵ is the **error term** (residuals)

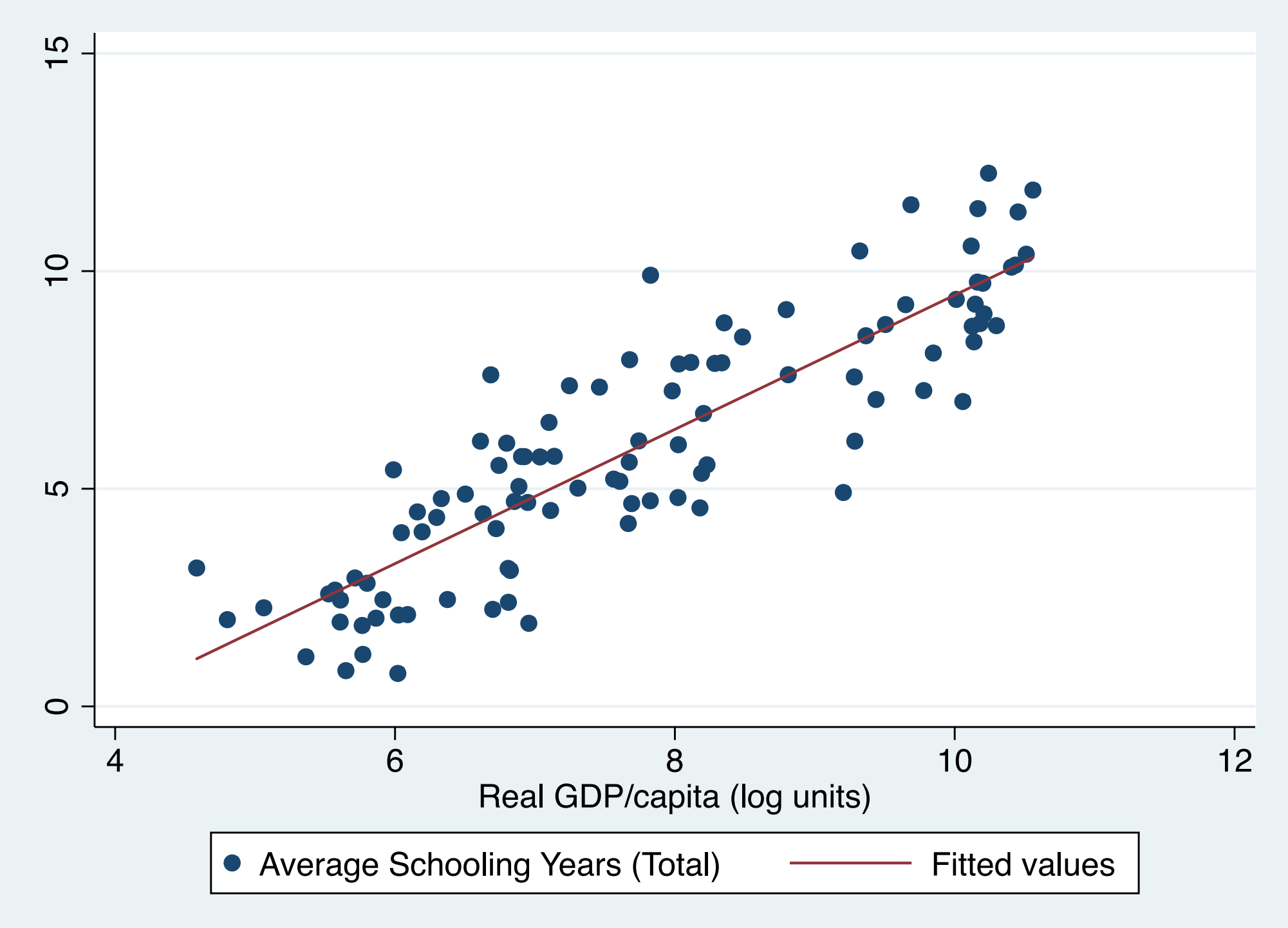
Note: the model assumes that the relationship is **linear** and **additive**.

The estimation of regression coefficients in a k-dimensional space is computationally more intensive, but is also based on least squares.


Fertility Rate (Births per Woman)
Average Schooling Years (Total)
Real GDP/capita (log units)
15
10
5
0
10
8
6
4
2
4
6
8
0
5
10
15


0
2
4
6
8
0
5
10
15
Average Schooling Years (Total)
Fertility Rate (Births per Woman)
Fitted values


0
2
4
6
8
4
6
8
10
12
Real GDP/capita (log units)
Fertility Rate (Births per Woman)
Fitted values


15
10
5
0
4
6
8
10
12
Real GDP/capita (log units)
Average Schooling Years (Total)
Fitted values

Multiple regression output: Unstandardised (metric)

`reg births schooling log_gdpc`

The `reg` command can take any number of **continuous** variables as arguments, and shows unstandardised coefficients by default, using their original metric and possible transformation:

```
. reg births schooling log_gdpc
```

Source	SS	df	MS
Model	150.301883	2	75.1509417
Residual	70.475313	83	.849100157
Total	220.777196	85	2.59737878

Number of obs	86
F(2, 83)	88.51
Prob > F	0.0000
R-squared	0.6808
Adj R-squared	0.6731
Root MSE	.92147

births	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
schooling	-.1976117	.0724595	-2.73	0.008	-.3417306	-.0534927
log_gdpc	-.4703416	.1324501	-3.55	0.001	-.7337796	-.2069036
_cons	7.950304	.6861182	11.59	0.000	6.585642	9.314965

Multiple regression output: Standardised (beta)

```
reg births schooling log_gdpc, beta
```

The `beta` option provides standardised "beta" coefficients, which normalize each variable to in the model to fit $\mathcal{D} \sim \mathcal{N}(0, 1)$ in order to provide coefficients with comparable **standard deviation** units:

births	Coef.	Std. Err.	t	P>|t|	Beta
schooling	-.1976117	.0724595	-2.73	0.008	-.3686479
log_gdpc	-.4703416	.1324501	-3.55	0.001	-.4800156
_cons	7.950304	.6861182	11.59	0.000	.

(identical output for overall model fit omitted)

Multiple regression output: Dummies

```
reg births schooling i.region
```

Categorical variables can be used as dummies, i.e. binary recodes of each category that are tested against a **reference category** to provide regression coefficients for the net effect of each category:

births	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
schooling	-.0415563	.0639718	-0.65	0.518	-.1688888	.0857763
log_gdpc	-.742187	.1380037	-5.38	0.000	-1.016876	-.4674975
region						
2	-.6523485	.5803126	-1.12	0.264	-1.807432	.5027349
3	.3682404	.254364	1.45	0.152	-.1380585	.8745393
4	1.411177	.2486027	5.68	0.000	.9163457	1.906008
5	1.167491	.337383	3.46	0.001	.4959471	1.839035
_cons	8.315004	.8006456	10.39	0.000	6.721359	9.908649

(identical output for overall model fit omitted)