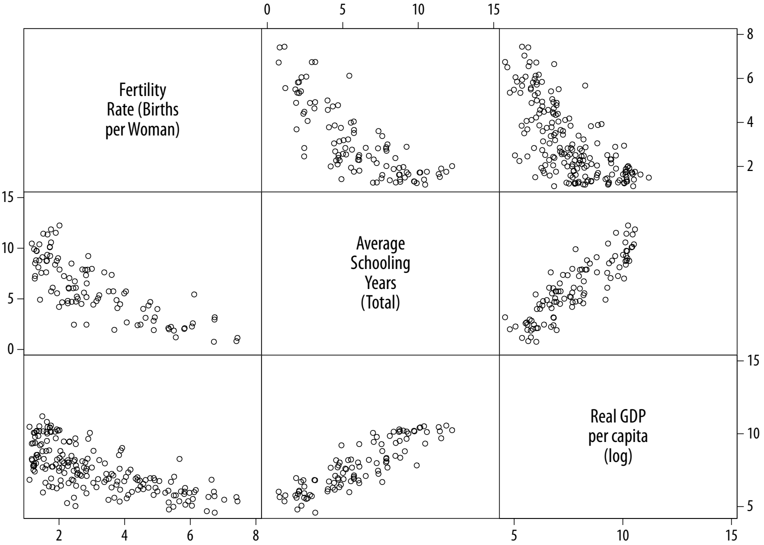
Linear regression


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

Multiple linear regression

```
. reg births schooling log_gdpc

      Source |       SS           df       MS      Number of obs   =        86
-------------+----------------------------------   F(2, 83)        =     88.51
       Model |  150.301883         2  75.1509417   Prob > F        =    0.0000
    Residual |   70.475313        83  .849100157   R-squared       =    0.6808
-------------+----------------------------------   Adj R-squared   =    0.6731
       Total |  220.777196        85  2.59737878   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 linear regression

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

Partial derivatives

Each coefficient is calculated by holding all others constant.

Least squares

The model is still optimized by minimizing the squared error terms.

Sanity check

The model is still assuming *linear*, *additive* relationships.

Logarithmic coefficients: see UCLA mini-guide

Linear-linear relationships: $Y = \alpha + \beta_1 X$

An increase of one unit of X is associated with an increase of β_1 units of Y.

Log-linear relationships: $\ln Y = \alpha + \beta_1 X$

An increase of one unit of X is associated with a $100 \times \beta_1\%$ increase in Y (true effect: $Y \times \exp(\beta_1)$).

Linear-log relationships: $Y = \alpha + \beta_1 \ln X$

A 1% increase in X is associated with a $0.01 \times \beta_1$ unit increase in Y (e.g. $\beta_1 \times \log(1.15)$ for +15% in X).

Log-log relationships: $\ln Y = \alpha + \beta_1 \ln X$

A 1% increase in X is associated with a $\beta_1\%$ increase in Y.

reg births schooling log_gdpc, beta

Each variable can be normalized to fit $\mathcal{D} \sim \mathcal{N}(0,1)$, so that their standardized coefficients have 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)

Sanity check

Interpret unstandardized coefficients; use standardization only for model comparisons.

Regression dummies and categorical predictors

Single coefficient of dummy X_3

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3(0) + \epsilon$$
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3(1) + \epsilon$$

The omitted category $X_3 = 0$ is called the reference category and is part of the baseline model $Y = \alpha$, for which all coefficients are null.

Example

$$\mathit{Income} = \alpha + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{education} + 0 \cdot \text{male} \qquad + \epsilon$$
$$\mathit{Income} = \alpha + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{education} + 1 \cdot \text{female} \qquad + \epsilon$$

reg births schooling log_gdpc 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)

Regression diagnostics

Residuals

- `predict yhat`: fitted values
- `predict r, resid`: residuals
- `predict r, rsta`: standardized residuals

Use `rvfplot` for residuals-versus-fitted values plots.

Heteroskedasticity

When the residuals are not normally distributed, the model expresses heterogeneous variance (unreliable standard errors).

Examples: UCLA *Regression with Stata*, ch. 2

There are many more diagnostics on display there.

Regression diagnostics

Interaction terms

- Use `vif` to detect variables with $VIF > 10$
- Use # and ## to capture interactions
- Use `c.` to interact continuous variables

(See also `avplot` and partial regression.)

Variance inflation

Variables that strongly interact induce multicollinearity ‘inside’ your model, making standard errors unreliable.

Examples: UCLA Stata FAQ

Search for “comparing coefficients across groups”.