Regression

Lecture 6

Regression

- We now assert that *X* causes *Y*, i.e. the growth rate influences the birth rate (not vice versa)
- Regression analysis determines the effect of *X* upon *Y*
- Regression analysis also allows several explanatory variables to influence *Y*

Terminology

- Regression equation: $Y=a+bX+e$
- Dependent variable is the variable we are trying to explain, Y
- We assert that Y is caused by the independent variable, X
- The model $Y=a+bX+e$ suggests that X affects Y in a linear way, with an intercept of value *a*, and slope *b*
- e is an error term, we can't predict Y perfectly just from knowing X

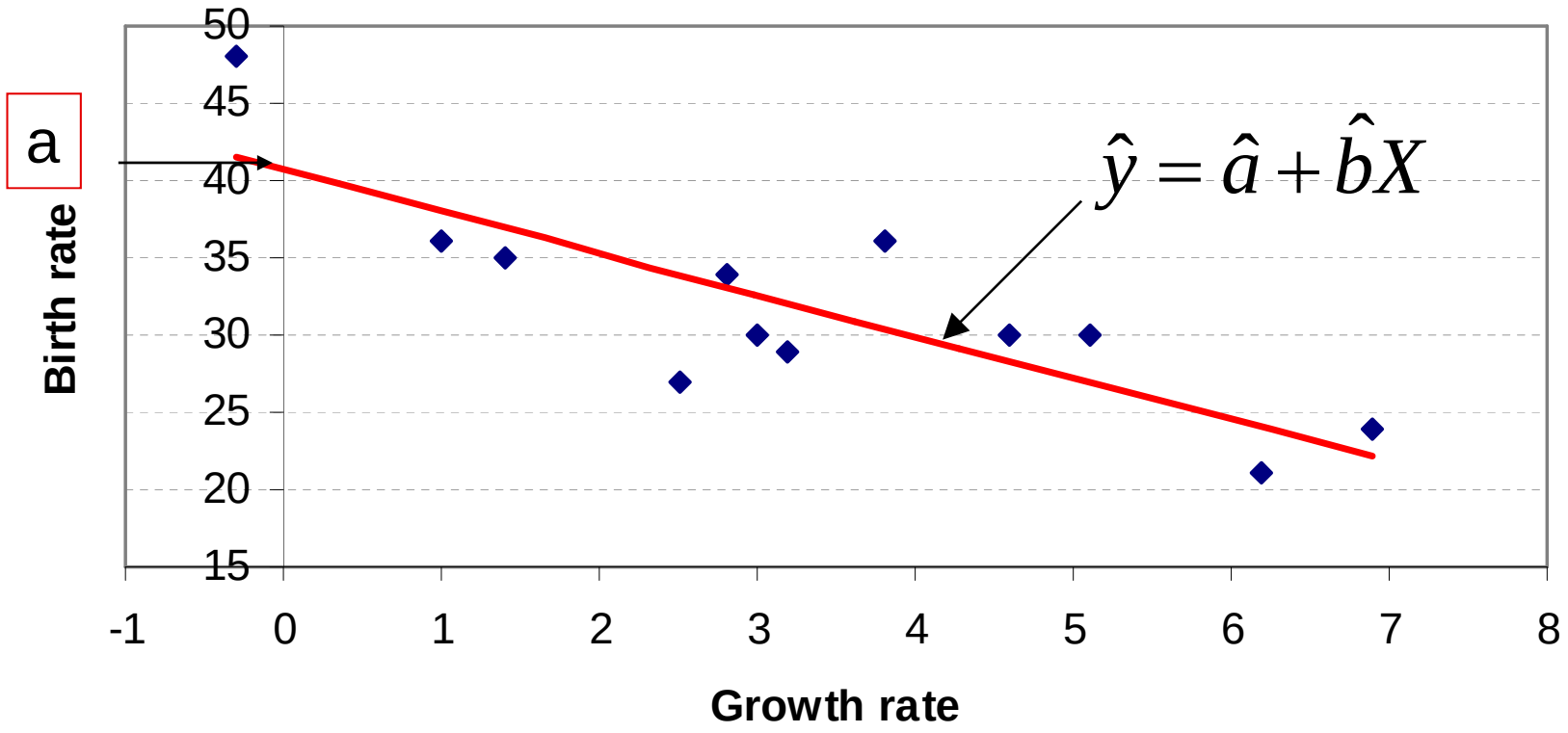
Some Notation

- *a* and *b* are the true population parameters of the relationship
- We estimate them with

$$\hat{y} = \hat{a} + \hat{b}X$$

- $\hat{y}$ will be our *estimate* of y. The difference between them is e, the error term

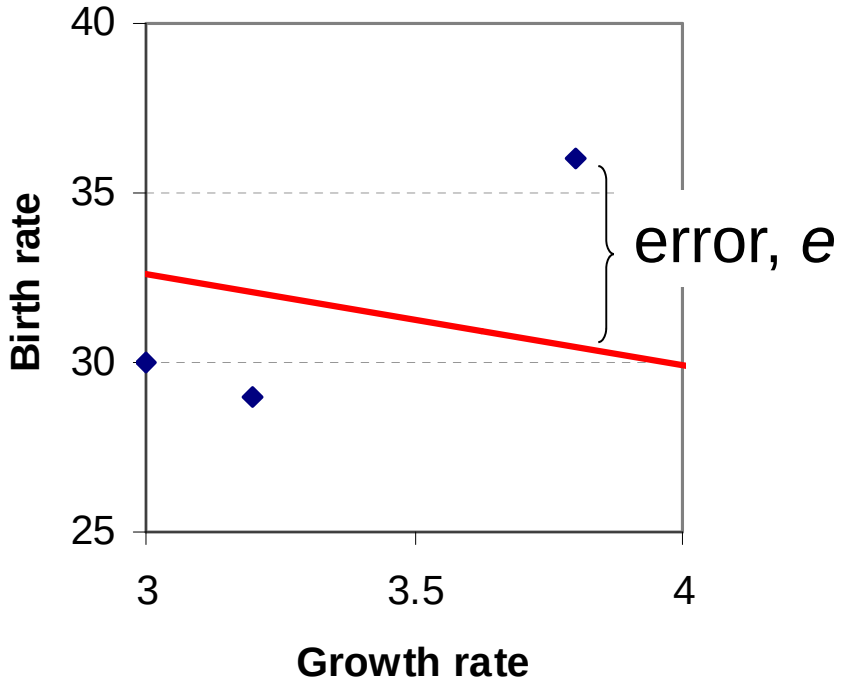
Regression Line: the Line of "Best Fit"



How to Obtain the Regression Line

- Minimise the sum of squared errors, ESS Σe^2

- The error is the vertical distance between an observation and the regression line



Regression Formulae

- Slope

$$\hat{b} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

- Intercept

$$\hat{a} = \bar{Y} - \hat{b}\bar{X}$$

Regression of Birth Rate on Growth Rate

Country	Birth rate Y	GNP growth X	Y^2	X^2	XY
Brazil	30	5.1	900	26.01	153.0
Colombia	29	3.2	841	10.24	92.8
Costa Rica	30	3.0	900	9.00	90.0
India	35	1.4	1,225	1.96	49.0
Mexico	36	3.8	1,296	14.44	136.8
Peru	36	1.0	1,296	1.00	36.0
Philippines	34	2.8	1,156	7.84	95.2
Senegal	48	–0.3	2,304	0.09	–14.4
South Korea	24	6.9	576	47.61	165.6
Sri Lanka	27	2.5	729	6.25	67.5
Taiwan	21	6.2	441	38.44	130.2
Thailand	30	4.6	900	21.16	138.0
Total	380	40.2	12,564	184.04	1,139.7

Calculation of Regression Coefficients

- Slope

$$\hat{b} = \frac{12 \times 1{,}139.70 - 40.2 \times 380}{12 \times 184.04 - 40.2^2} = -2.700$$

- Intercept

$$\hat{a} = \frac{380}{12} - (-2.700) \times \frac{40.2}{12} = 40.711$$

- $Y_i = 40.71 - 2.70X_i + e_i$

$$\hat{Y}_i = 40.71 - 2.70X_i$$

Interpretation of the Coefficients

$$\hat{Y}_i = 40.71 - 2.70 X_i$$

- *a* is the intercept - when growth is zero, birth rate is predicted to be 40.71
- *b* is the slope - an extra 1 percentage point in the growth rate leads to a *fall* in the birth rate of 2.70 births per thousand

Prediction

- The regression line may be used for prediction
- To predict the birth rate for a country growing at 3% p.a. we insert this value into the regression equation

$$\hat{Y} = 40.71 - 2.7 \times 3 = 32.6$$

- The predicted birth rate is 32.6

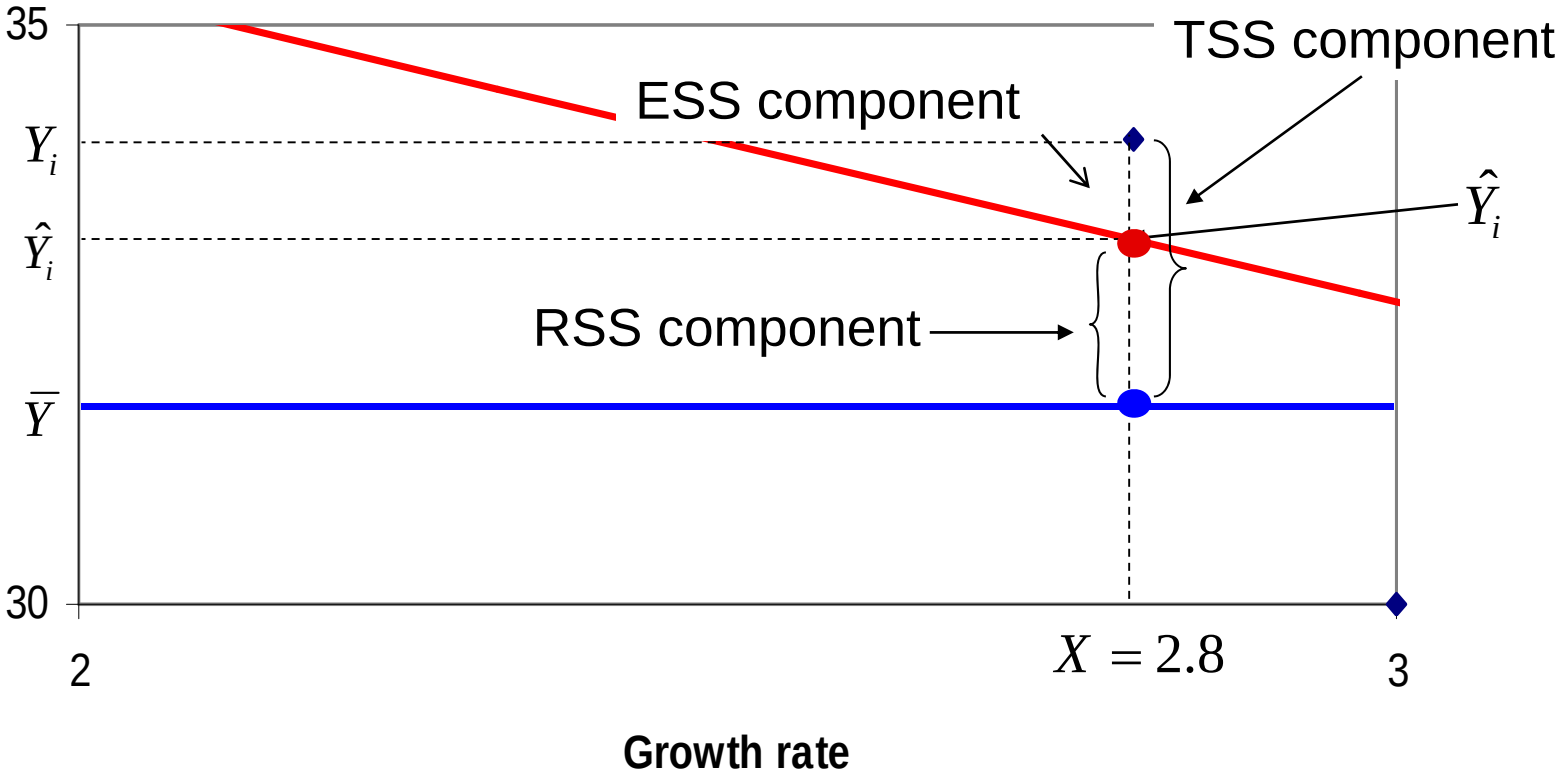
Measuring the Goodness of Fit

- Use the coefficient of determination, R^2

$$R^2 = \frac{RSS}{TSS}$$

- $0 \leq R^2 \leq 1$
- RSS: regression sum of squares
  TSS: total sum of squares
- NB: notation varies in different textbooks

The Component Parts of R^2



Calculating Sums of Squares

- TSS = ESS + RSS

$$TSS = \sum\left(Y - \bar{Y}\right)^2 = \sum Y^2 - n\bar{Y}^2 = 12,564 - 12 \times 31.67^2 = 530.67$$

$$ESS = \sum\left(Y - \hat{Y}\right)^2 = \sum Y^2 - a\sum Y - b\sum XY$$

$$= 12,564 - 40.71 \times 380 - \left(-2.7\right) \times 1,139.7 = 170.75$$

$$RSS = 530.67 - 170.75 = 359.92$$

- Hence $R^2 = \dfrac{359.92}{530.67} = 0.678$
  - Our model explains 67.8% of the variation in birth rates
  - Note that $r^2 = R^2$ in the case of 1 independent var

Regression Example

Birth-rate=*a*+*b* growth-rate+e

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.823546
R Square	0.678227
Adjusted R Square	0.64605
Standard Error	4.132239
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	359.9127	359.9	21.08	0.000994
Residual	10	170.753967	17.08		
Total	11	530.666667			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	40.71173	2.30313463	17.68	7E-09	35.58003	45.84344
Growth rate	-2.70002	0.5881036	-4.591	1E-03	-4.0104	-1.389644

Hypothesis Testing

- Often we want to test statistical significance of $\hat{b}$
- H_0: b=0; H_1: b≠0

$$t = \frac{\hat{b} - b}{s.e.(\hat{b})} \sim t_{n-2}$$

- In the example t=-4.59, so we can reject the null and conclude that growth does have a significant effect on birth rates

Testing the Goodness of Fit

- Is the model any good? Is it better than Birth-rate=*a*+*e*?
- H_0: *b*=0

$$F = \frac{R^2 / k}{(1 - R^2)/(n - k - 1)} \sim F_{k, n-k-1, \alpha}$$

- in the example, R^2=0.678, F=21.08
- more useful when we have a multiple regression model, such as
  - Birth-rate=*a*+*b*(growth-rate)+*c*(health-expenditure) +*d*(GDP) +*e*

Summary

- Correlation measures the association between two variables
- Regression extends this by
  - measuring the effect of *X* upon *Y* (the slope coefficient *b*)
  - Allowing us to incorporate more explanatory variables
- The regression line is found by minimizing the sum of the squared errors. This is the 'line of best fit'
- A measure of goodness of fit is the R^2