Basic Biostatistics

Statistics for Public Health Practice

B. Burt Gerstman

Chapter 14: Correlation and Regression

4/4/2010 1

In Chapter 14:

4/4/2010 2

Data

- Quantitative explanatory variable *X*
- Quantitative response variable *Y*
- Objective: To quantify the *linear* relationship between X and Y

Table 14.1 Synonyms for *explanatory variable* and *response variable*.

Explanatory Variable	→	Response Variable
X	→	*Y*
independent variable	→	dependent variable
factor	→	outcome
treatment	→	response
exposure	→	disease

4/4/2010 3

Illustrative Data (Doll, 1955)

per capita cigarette consumption (X)

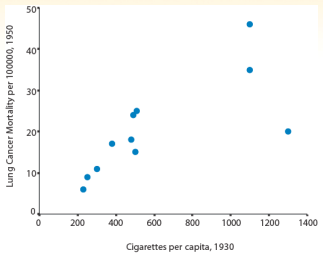
lung cancer mortality per 100,000 in 1950 (Y)

COUNTRY	CIG1930	LUNGCA
USA	1300	20
Great Britain	1100	46
Finland	1100	35
Switzerland	510	25
Canada	500	15
Holland	490	24
Australia	480	18
Denmark	380	17
Sweden	300	11
Norway	250	9
Iceland	230	6

$n = 11$

4/4/2010 4

Scatterplot

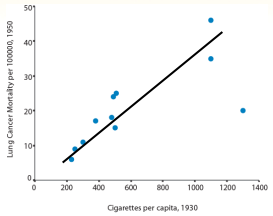


Assess:

- Form
- Direction of association
- Outliers
- Strength of relation

4/4/2010 5

Doll, 1955

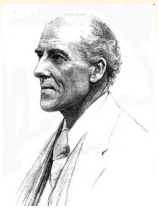


- Form: linear
- Direction: positive association
- Outlier: no clear outliers
- Strength: difficult to determine by eye

4/4/2010 6

Correlation Coefficient *r*

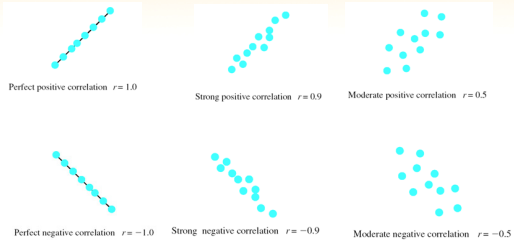
- *r* ≡ *Pearson's product-moment correlation coefficient*
- Measures degree to which *X* and *Y* "go together"
- Always between –1 and 1
- r ≈ 0 ⇒ no correlation
- *r* > 0 ⇒ positive correlation
- *r* < 0 ⇒ negative correlation
- Closer *r* is to 1 or –1, the **stronger** the correlation



Karl Pearson
1857 - 1936

4/4/2010 8

Correlational Direction and Strength



4/4/2010 9

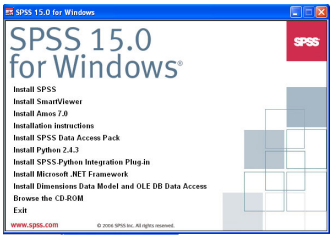
Interpretation of *r*

- **Direction of association:** positive, negative, ~0
- **Strength of association**
  - close to 1 or –1 ⇒ "strong"
  - close to 0 ⇒ "weak"
  - guidelines
    - if |*r*| ≥ .7 ⇒ say "strong"
    - if |*r*| ≤ .3 ⇒ say "weak"

4/4/2010 10

Calculating *r*

By hand, calculator or computer program
We opt for latter



4 13

SPSS output

SPSS > Analyze > Correlate > Bivariate

Correlations

		Cig Consuption per capita, 1930	Lung Cancer Mortality per 100000, 1950
Cig Consuption per capita, 1930	Pearson Correlation	1	.737**
	Sig. (2-tailed)		.010
	N	11	11
Lung Cancer Mortality per 100000, 1950	Pearson Correlation	.737**	1
	Sig. (2-tailed)	.010	
	N	11	11

**. Correlation is significant at the 0.01 level (2-tailed).

r

r = 0.74 indicates a strong, positive association

4/4/2010 14

Coefficient of determination (r^2)

- Square the correlation coefficient ⇒ r^2 = proportion of variance in Y mathematically explained by X
- Illustrative data: $r^2 = 0.737^2 = 0.54$ ⇒ 54% of variance in lung cancer mortality is mathematically explained per capita smoking rates

4/4/2010 15

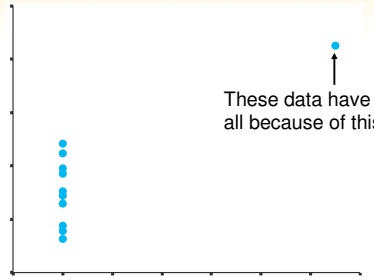
Cautions

- Outliers
- Non-linear relations
- Confounding (correlation is NOT causation)
- Randomness

4/4/2010 16

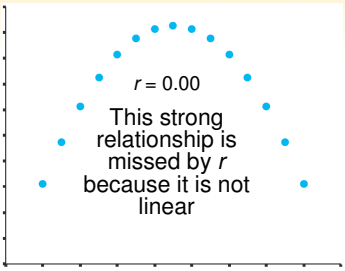
Outliers

Outliers can have profound influence on *r*



4/4/2010 17

Linear Relations Only

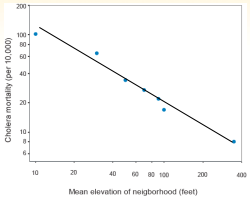


4/4/2010 18

Confounding

Correlation ≠ Causation

William Farr showed this strong negative correlation between cholera mortality and elevation above sea level in defense of miasma theory

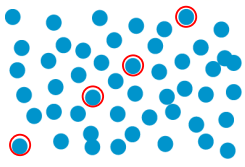


However, he failed to account for the fact that people who lived at low elevations were more likely to drink from contaminated water sources (∴ confounding)

4/4/2010 19

Don't be fooled by randomness

Selection of specific data points would result in a false correlation



4/4/2010 20

Hypothesis Test

Test the claim H_0: ρ = 0
where ρ ≡ correlation coefficient parameter

SPSS > Analyze > Correlate > Bivariate output:

Correlations

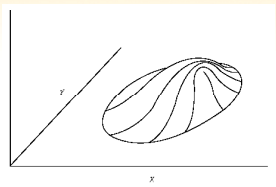
		Cig Consuption per capita, 1930	Lung Cancer Mortality per 100000, 1950
Cig Consuption per capita, 1930	Pearson Correlation	1	.737**
	Sig. (2-tailed)		.010
	N	11	11
Lung Cancer Mortality per 100000, 1950	Pearson Correlation	.737**	1
	Sig. (2-tailed)	.010	
	N	11	11

**. Correlation is significant at the 0.01 level (2-tailed).

∴ P = .010 (two-sided) ⇒ reliable evidence against H_0
⇒ the correlation is statistically significant

4/4/2010 23

Bivariate Normality



Strictly speaking: *P*-value requires Normality of the joint distribution of X and Y ("bivariate Normality")

§14.4. Regression

Regression model (equation for line):

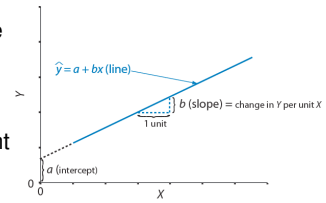
$$\hat{y}_i = a + b \cdot X_i$$

where

$\hat{y}_i \equiv$ predicted value of Y at x_i

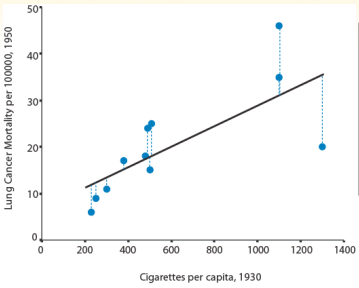
$a \equiv$ intercept coefficient

$b =$ slope coefficient



Least Squares Line

Residual ≡ distance of data point from regression line (dotted)



The best fitting line minimizes the residuals

Determine a and b of best fitting line via formula, calculator, or computer.

Coefficient by SPSS

Analyze > Regression > Linear

Slope estimate (*b*)

Intercept estimate (*a*)

Coefficients[a]

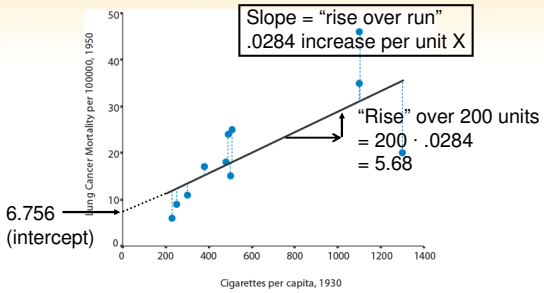
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	6.756	4.906		1.377	.202
	CIG1930	2.284E-02	.007	.737	3.275	.010

a. Dependent Variable: MORTALIT

Regression line:

$\hat{y} = 6.756 + 0.0284 \cdot X$

$\hat{y} = 6.756 + 0.0284 \cdot X$



Population Regression Model

$$y_i = \alpha + \beta x_i + \epsilon_i$$

where

- $\alpha \equiv$ intercept parameter
- $\beta \equiv$ slope parameter
- $\epsilon_i \equiv$ residual error, observation *i*

Objective:
To estimate β with (1 – α)100% confidence

CI for β

Analyze > Regression > Linear > Statistics

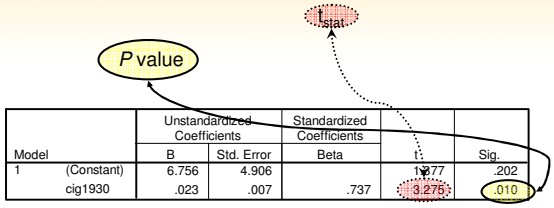
SPSS statistics options Dialogue box

95% CI for β

Model		95% Confidence Interval for B	
		Lower Bound	Upper Bound
1	(Constant)	-4.342	17.854
	cig1930	.007	.039

95% CI for β (.007 to.039)

Testing H_0: β = 0



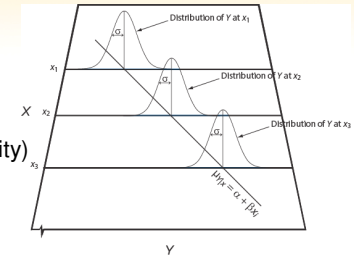
$df = n - 2 = 11 - 2 = 9$

$\therefore P = .010 \Rightarrow$ evidence against H_0 is good

$\Rightarrow$ the slope is statistically significant

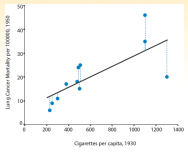
Conditions for Regression Inference

- Linearity
- Independent observations
- Normality
- Equal variance (homoscedasticity)



Assessing L.I.N.E

- Inspect scatterplot for linearity
- Inspect residuals for
  - linearity
  - Normality
  - equal variance



i	COUNTRY	X	Y	predicted $\hat{y}$	residual $y_i - \hat{y}_i$
1	USA	1300	20	36.453	−16.453
2	GrBritain	1100	46	31.884	14.116
3	Finland	1100	35	31.884	3.116
4	Switzerland	510	25	18.406	6.594
5	Canada	500	15	18.178	−3.178
6	Holland	490	24	17.950	6.050
7	Australia	480	18	17.721	0.279
8	Denmark	380	17	15.437	1.563
9	Sweden	300	11	13.609	−2.609
10	Norway	250	9	12.467	−3.467
11	Iceland	230	6	12.010	−6.010

Assessing Conditions

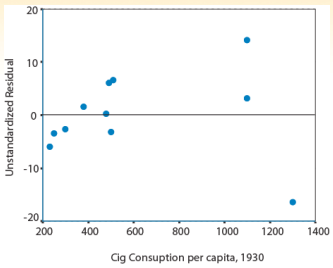
residual $y_i - \hat{y}_i$
−16.453
14.116
3.116
6.594
−3.178
6.050
0.279
1.563
−2.609
−3.467
−6.010

$\Rightarrow$

```
-1|6
-0|2336
 0|01366
 1|4
x10
```

$\Rightarrow$ no major departures from Normality

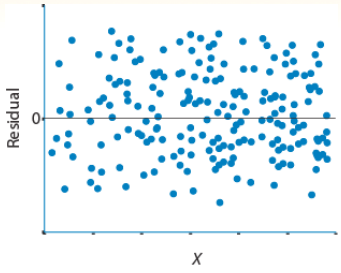
Residual plotted against X values



Data too sparse to assess

Residual Plot

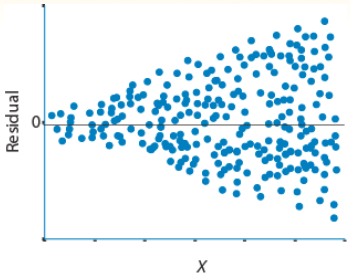
Example of linearity with equal variance



4/4/2010 43

Residual Plot

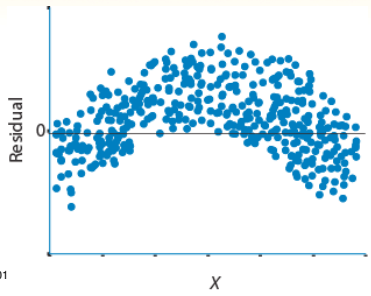
Example of linearity with unequal variance



4/4/2010 44

Residual Plot

Example of non-linearity with equal variance



4/4/201 45