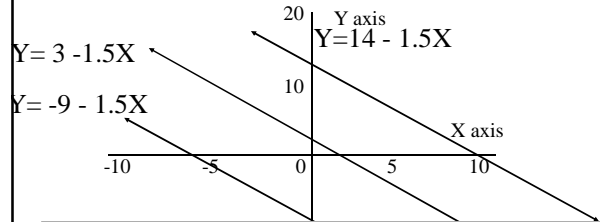




Linear Regression

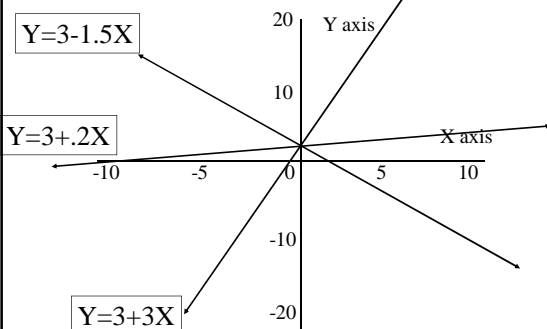
Review: $Y = a + bX$



- ◆ The “constant” or “intercept” (a)
 - Determines where the line intersects the Y-axis
 - If a increases (decreases), the line moves up (down)

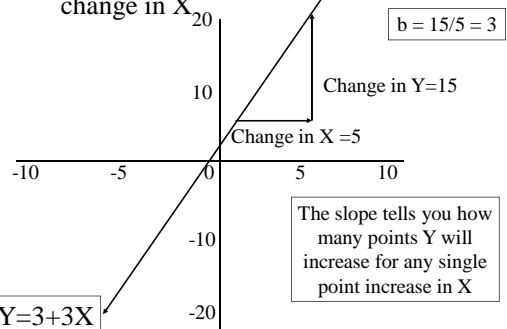
Review: $Y = a + bX$

- ◆ The slope (b) determines the steepness of the line




Review: Slopes

- ◆ The slope (b) is the ratio of change in Y to change in X



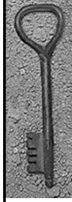
The slope tells you how many points Y will increase for any single point increase in X



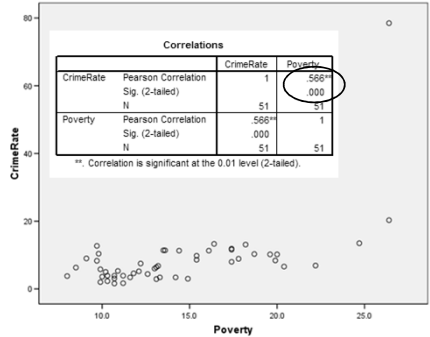
Regression: Example

- ◆ Y: Crime Rate
- ◆ X: Poverty Rate
- ◆ Hypothesis:
 - There is an association between the two variables
 - High poverty rate causes high crime rate

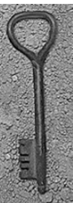
GOG 502/PLN 504 Youqin Huang 5



Example: Poverty and Crime Rate



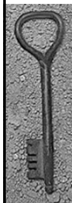
GOG 502/PLN 504 Youqin Huang 6



Hypothesis Test on r

- ◆ Is observed (positive or negative) correlation significantly different from zero?
 - Might the population have no linear association?
 - Population correlation denoted by greek “r”, rho (ρ)
- ◆ H0: There is no linear association ($\rho = 0$)
- ◆ H1: There is linear association ($\rho \neq 0$)

GOG 502/PLN 504 Youqin Huang 7



Correlation Coefficient (r)

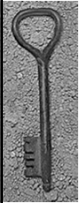
| Correlations | | | |
|--------------|---------------------|-----------|---------|
| | | CrimeRate | Poverty |
| CrimeRate | Pearson Correlation | 1 | .566** |
| | Sig. (2-tailed) | | .000 |
| | N | 51 | 51 |
| Poverty | Pearson Correlation | .566** | 1 |
| | Sig. (2-tailed) | .000 | |
| | N | 51 | 51 |

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

p-value: The probability of observing r if $\rho = 0$.

Compare it to α !

GOG 502/PLN 504 Youqin Huang 8



Example

- ◆ So there is a significant linear association
- ◆ What form of linear association is there between crime and poverty rate?
 - Linear Regression
 - Y vs. X: theory determines the direction of causation;

GOG 502/PLN 504 Youqin Huang

9

| State | ViolentRate | CrimeRate | Poverty |
|-------|-------------|-----------|---------|
| AK | 761 | 9 | 11.5 |
| VT | 750 | 12 | 17.4 |
| AR | 693 | 10 | 20.0 |
| AZ | 627 | 5 | 10.7 |
| CA | 627 | 5 | 10.7 |
| CO | 627 | 5 | 10.7 |
| CT | 627 | 5 | 10.7 |
| DE | 627 | 5 | 10.7 |
| FL | 627 | 5 | 10.7 |
| GA | 627 | 5 | 10.7 |
| HI | 627 | 5 | 10.7 |
| IA | 627 | 5 | 10.7 |
| ID | 627 | 5 | 10.7 |
| IL | 627 | 5 | 10.7 |
| IN | 627 | 5 | 10.7 |
| KS | 627 | 5 | 10.7 |
| KY | 627 | 5 | 10.7 |
| LA | 627 | 5 | 10.7 |
| MA | 627 | 5 | 10.7 |
| MD | 627 | 5 | 10.7 |
| ME | 627 | 5 | 10.7 |
| MI | 627 | 5 | 10.7 |
| MN | 627 | 5 | 10.7 |
| MO | 627 | 5 | 10.7 |
| MS | 627 | 5 | 10.7 |
| MT | 627 | 5 | 10.7 |
| NC | 627 | 5 | 10.7 |
| ND | 627 | 5 | 10.7 |
| RI | 627 | 5 | 10.7 |
| SC | 627 | 5 | 10.7 |
| SD | 627 | 5 | 10.7 |
| TN | 627 | 5 | 10.7 |
| TX | 627 | 5 | 10.7 |
| VA | 627 | 5 | 10.7 |
| WV | 627 | 5 | 10.7 |
| DC | 627 | 5 | 10.7 |

Theory determines direction of causation

SPSS Output

| Coefficients ^a | | | | | |
|---------------------------|------------|-----------------------------|---------------------------|--------|------|
| Model | | Unstandardized Coefficients | Standardized Coefficients | t | Sig. |
| 1 | (Constant) | -10.136 | | -2.460 | .017 |
| | Poverty | 1.323 | .275 | 4.804 | .000 |

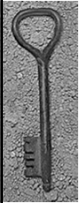
a. Dependent Variable: CrimeRate

- ◆ a = constant = -10.136
- ◆ b = slope = parameter estimate for ind. var. = 1.323
- ◆ So the equation is:

$$\text{crimeRate} = -10.136 + 1.323 * \text{poverty}$$
 One more percentage of poverty rate adds 1.323 percentage in crime rate

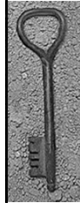
GOG 502/PLN 504 Youqin Huang

12



Hypothesis Tests: Slopes

- ◆ Given: Observed slope relating poverty to crime rate is = 1.32
- ◆ Question: Can we generalize this to the population? Is it significantly different from zero?
 - After all, every sample yields a different value of b
 - Is this “positive” slope merely the product of chance?
- ◆ Notation: slope = b, population slope = β
- ◆ H0: Population slope $\beta = 0$
- ◆ H1: Population slope $\beta \neq 0$
- ◆ Or one-tailed test: H0: $\beta \leq 0$, H1: $\beta > 0$

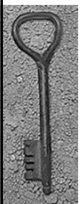


Hypothesis Tests: Slopes

- ◆ The sampling distribution of the slope (b) approximates a T-distribution

$$t_{N-2} = \frac{b_{YX}}{s_b} = \frac{b_{YX}}{\sqrt{\frac{MS_{ERROR}}{s_X^2(N-1)}}}$$

- Where s_b is the sample point estimate of the standard error
- The t-value is based on N-2 degrees of freedom



Hypothesis Tests: Constant

- ◆ You can also use a T-test to determine if the constant (a) is significantly different from zero
- Hypotheses (α = population parameter of a):
- H0: $\alpha = 0$, H1: $\alpha \neq 0$

$$t_{N-2} = \frac{a_{YX}}{\sqrt{\frac{MS_{ERROR}}{(N-1)}}}$$

- But, most research focuses on slopes



Hypothesis Test: constant, slope


| Coefficients ^a | | | | | | |
|---------------------------|------------|-----------------------------|------------|---------------------------|--------|------|
| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | -10.138 | 4.121 | | -2.460 | .017 |
| | Poverty | 1.323 | .275 | .566 | 4.804 | .000 |

a. Dependent Variable: CrimeRate

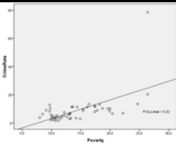
- H0: $\alpha = 0$
- H1: $\alpha \neq 0$

What can we conclude?

- ◆ H0: Population slope $\beta = 0$
- ◆ H1: Population slope $\beta \neq 0$

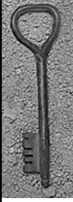


R-Square



- ◆ Issue: Even the “best” regression line misses data points. We still have some errors.
- ◆ Q: How good is our line at summarizing the relationship between two variables?
 - Do we have a lot of error? Or only a little? (i.e., the line closely estimates cases)
- ◆ Solution: The R-Square statistic

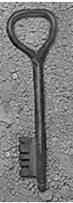
GOG 502/PLN 504 Youqin Huang 17



Properties of R-Square

- ◆ Ranges from 0 to 1
 - 1 indicates that perfect prediction of Y by X
 - 0 indicates that the line explains no variance in Y
- ◆ Tells us the proportion of all variance in Y that is explained as a linear function of X
 - It measures “how good” our line is at predicting Y (goodness of fit)
- ◆ The R-square indicates how well variables (or groups of variables) account for variation in Y (explanatory power)

GOG 502/PLN 504 Youqin Huang 18



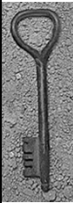
Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1 | .566 ^a | .320 | .306 | 8.926 |

a. Predictors: (Constant), Poverty

R-square=0.32: 32% of variation in crime rate is explained by poverty rate

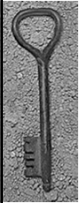
GOG 502/PLN 504 Youqin Huang 19



Interpreting R-Square

- ◆ R-square is often used as an overall indicator of the “success” of a regression model
 - Higher R-square is considered “better” than lower
 - But, don’t get over-zealous at increasing R-square
 - Not all variables that generate high R-square are sensible to include in a regression analysis
- ◆ Adjusted R-square: $Adjusted R^2 = 1 - \frac{n-1}{n-k} (1 - R^2)$
 - adjust R-square by degrees of freedom
 - k=number of parameters in the model: intercept + indep vars.
 - “adjusted R-square” is more conservative.

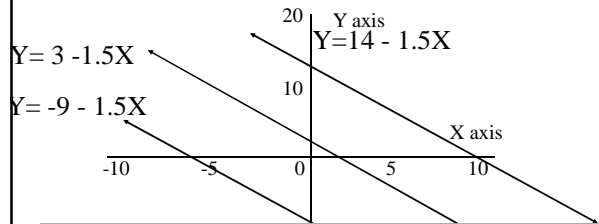
GOG 502/PLN 504 Youqin Huang 20



Questions:

- ◆ But how does SPSS (or other software) choose a, b to best describe the data (the best model)?
- ◆ How does SPSS calculate R-square?

Lines: $Y = a + bX$



- ◆ The “constant” or “intercept” (a)
 - Determines where the line intersects the Y-axis
 - If a increases (decreases), the line moves up (down)



The Linear Regression Model

- ◆ To model real data, we must take into account that points will miss the line
 - The deviation of points from the estimated value is called “error” (e_i)
- ◆ In regression the estimated value is derived from the formula $Y = a + bX$
 - Estimation is based on the value of X, slope, and constant (assumes linear relationship between X and Y)



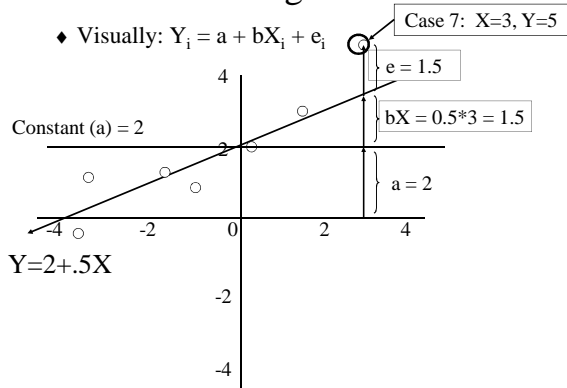
The Linear Regression Model

- ◆ The value of any point (Y_i) can be modeled as:

$$Y_i = a + b X_i + e_i$$
- The value of Y for case (i) is made up of
 - A constant (a)
 - A sloping function of the case’s value on variable X (b)
 - An error term (e), the deviation from the line; also called “residual”
- By adding error (e), an abstract mathematical function can be applied to real data points

The Linear Regression Model

◆ Visually: $Y_i = a + bX_i + e_i$



Estimating Linear Equations

◆ Q: How do we choose the best line to describe our real data?

◆ Answer: Look at the error

- If a given line formula misses points by a lot, the observed error will be large
- If the line is as close to all points as possible, observed error will be small

◆ Of course, even the best line has some error

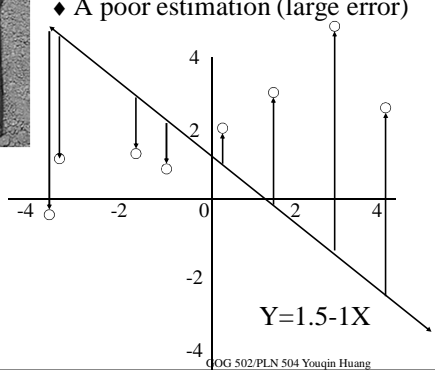
- Except when all data points are perfectly on a line

GOG 502/PLN 504 Youqin Huang

26

Estimating Linear Equations

◆ A poor estimation (large error)

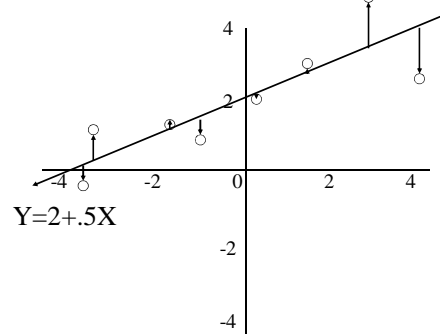


GOG 502/PLN 504 Youqin Huang

27

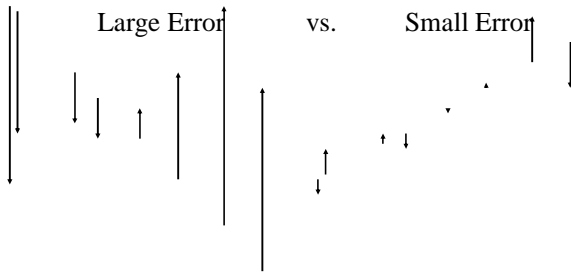
Estimating Linear Equations

◆ Better estimation (small error)



Estimating Linear Equations

- ◆ Look at the improvement (reduction) in error:



Estimating Linear Equations

- ◆ Idea: The “best” line is the one that has the least error (deviation from the line)
- ◆ Total deviation from the line can be expressed as:

$$\sum_{i=1}^N (Y_i - \hat{Y}_i) = \sum_{i=1}^N e_i$$

- To make all deviation positive, we square it, producing the “sum of squared error”

$$\sum_{i=1}^N (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^N e_i^2$$

GOG 502/PLN 504 Youqin Huang

30

Estimating Linear Equations

- ◆ Goal: Find values of constant (a) and slope (b) that produce the lowest squared error
 - Least Squares Regression
- ◆ The formula for the slope (b) that yields the “least squares error” is:

$$b = \frac{s_{YX}}{s_X^2}$$

- Where s_X^2 is the variance of X
- And s_{YX} is the covariance of Y and X

GOG 502/PLN 504 Youqin Huang

31

Covariance

- ◆ Variance: Sum of deviation about Y-bar squared

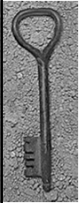
$$s_Y^2 = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{N-1} = \frac{\sum_{i=1}^N (Y_i - \bar{Y})(Y_i - \bar{Y})}{N-1}$$

- Covariance (s_{YX}): Sum of deviation about Y-bar multiplied by deviation around X-bar:

$$s_{YX} = \frac{\sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X})}{N-1}$$

GOG 502/PLN 504 Youqin Huang

32



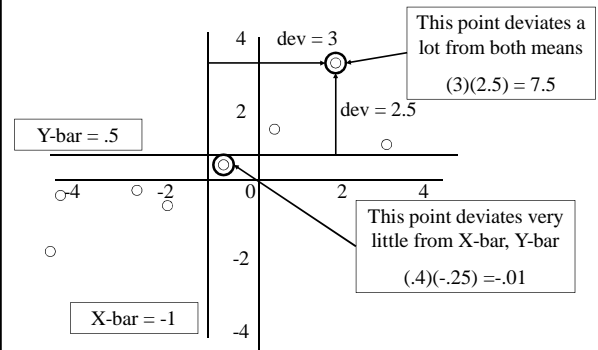
Covariance

- ◆ Covariance: A measure of how much variance of a case in X is accompanied by variance in Y
- ◆ It measures whether deviation (from mean) in X tends to be accompanied by similar deviation in Y
 - Or if cases with positive deviation in X have negative deviation in Y
 - This is summed up for all cases in the data
- ◆ The covariance is another numerical measure that characterizes the extent of linear association
 - As is the correlation coefficient (r)

GOG 502/PLN 504 Youqin Huang

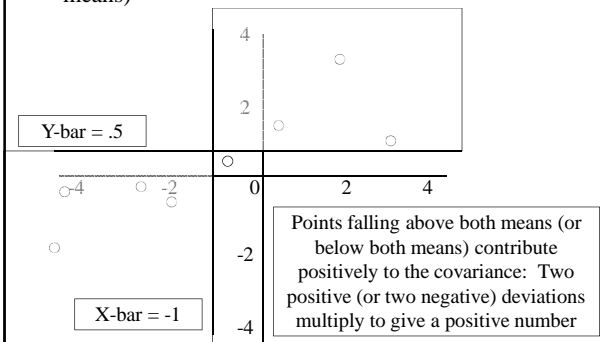
33

Covariance



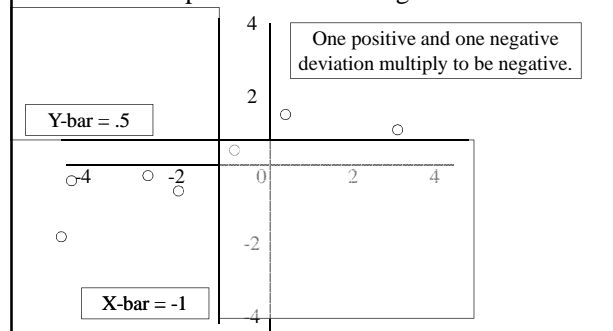
Covariance


- ◆ Some points fall above both means (or below both means)



Covariance

- ◆ Points falling above one mean but below the other = one positive and one negative deviation

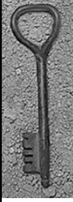




Properties of Covariance

- ◆ If positive deviation on X tends to be accompanied by positive deviation in Y (or negative X deviation with negative Y):
 - Then the covariance sums to a large positive number
- ◆ If positive X deviation is accompanied by negative deviation on Y (or vice versa)
 - Then the covariance sums to a large negative number
- ◆ If points are scattered all around, positives and negatives cancel out – the covariance is near zero

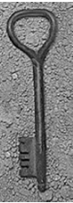
GOG 502/PLN 504 Youqin Huang 37



Covariance and Slope

- ◆ The covariance has properties similar to the slope
 - If points cluster from lower-left to upper right, the slope & covariance are positive
 - And upper left to lower right = negative
- ◆ The covariance can be used to calculate a regression slope that minimizes error for all points
 - The “Ordinary Least Squares” (OLS) error slope

GOG 502/PLN 504 Youqin Huang 38



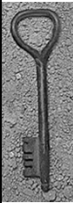
Covariance and Slope

- ◆ The slope formula can be written out as follows:

$$b = \frac{s_{YX}}{s_X^2}$$

$$b = \frac{\frac{\sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X})}{N-1}}{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N-1}} = \frac{\sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^N (X_i - \bar{X})^2}$$

GOG 502/PLN 504 Youqin Huang 39



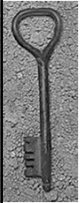
Computing the Constant

- ◆ Once the slope has been calculated, it is simple to determine the constant (a):

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

- Simply plug in the values of Y-bar, X-bar, and b
- The calculated value of b is called a “coefficient”
- The value of a is called the constant or intercept

GOG 502/PLN 504 Youqin Huang 40



Computing Regressions

- ◆ Regression coefficients can be calculated in statistical software (e.g. SAS, SPSS)
 - You will rarely, if ever, do them by hand
- ◆ Statistical software will also estimate:
 - The constant (a)
 - Related statistics and results of hypothesis testing procedures



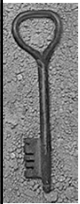
SPSS Output

| Coefficients ^a | | | | | | |
|---------------------------|------------|-----------------------------|------------|---------------------------|--------|------|
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | -10.136 | 4.121 | | -2.460 | .017 |
| | Poverty | 1.323 | .275 | .566 | 4.804 | .000 |

a. Dependent Variable: CrimeRate

- ◆ a = constant = -10.136
- ◆ b = slope = parameter estimate for ind. var. = 1.323
- ◆ So the equation is:

$$\text{crimeRate} = -10.136 + 1.323 * \text{poverty}$$
 One more percentage of poverty rate adds 1.323 percentage in crime rate

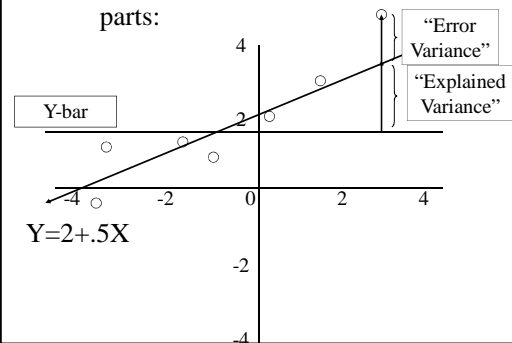



R-Square

- ◆ Issue: Even the “best” regression line misses data points. We still have some errors.
- ◆ Q: How good is our line at summarizing the relationship between two variables?
 - Do we have a lot of error? Or only a little? (i.e., the line closely estimates cases)
- ◆ Solution: The R-Square statistic

R-Square

- ◆ Variance around Y-bar can be split into two parts:



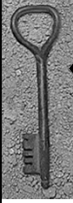


R-Square

- The total variation of a case Y_i around \bar{Y} can be partitioned into two parts (like ANOVA):
- 1. Explained variance
 - Also called "Regression variance", "Model variance"
 - The variance we predicted based on the line
- 2. Error variance, or Residual
 - The variance not accounted for by the line
- Summing squared deviation for all cases give us:

$$SS_{TOTAL} = SS_{REGRESSION} + SS_{ERROR}$$

GOG 502/PLN 504 Youqin Huang 45



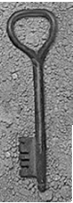
R-Square

- The R-Square statistic is computed as follows:

$$R^2_{YX} = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{SS_{REGRESSION}}{SS_{TOTAL}} = \frac{s^2_{YX}}{s^2_X s^2_Y}$$

- Q: What is R-square if the line is perfect?
- Answer: R-square = 1.00
- Q: What is R-square if the line is NO HELP in estimating points... (lots of error)
- Answer: R-square is zero

GOG 502/PLN 504 Youqin Huang 46



| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|-------|-------------------|----------|-------------------|----------------------------|
| 1 | .566 ^a | .320 | .306 | 6.926 |

a. Predictors: (Constant), Poverty

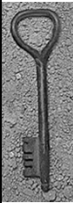
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|----|-------------|--------|-------------------|
| 1 | Regression | 1839.069 | 1 | 1839.069 | 23.081 | .000 ^b |
| | Residual | 3904.252 | 49 | 79.679 | | |
| | Total | 5743.322 | 50 | | | |

a. Predictors: (Constant), Poverty
b. Dependent Variable: CrimeRate

The R-Square indicate how well the line summarizes the data, or how much of the variance can be explained by the model.

$$R^2 = \frac{SS_{REGRESSION}}{SS_{TOTAL}} = \frac{1839.069}{5743.321} = 0.3202$$

GOG 502/PLN 504 Youqin Huang 47



R-Square & Correlation Coefficient

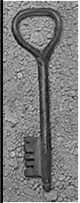
- The R-square is literally the square of r

$$r = \frac{s_{YX}}{s_X s_Y}$$

$$R^2_{YX} = \frac{s^2_{YX}}{s^2_X s^2_Y}$$

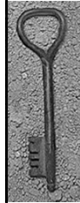
- r is a measure of linear association
- Ranges from -1 to 1
- 0 = no linear association
- 1 = perfect positive linear association
- 1 = perfect negative linear association
- R²: share of explained variance
- Ranges from 0 to 1
- 0 = no association
- 1 = perfect positive linear association

GOG 502/PLN 504 Youqin Huang 48



Covariance, R-square, r, and b

- ◆ All provide information about the relationship between X and Y
- ◆ Covariance and r can be positive or negative
 - r is scaled from -1 to +1, covariance is not
- ◆ b tells you the actual slope
 - It relates change in X to change in Y in real units
- ◆ R-square is like r, but is never negative
 - it tells you “explained” variance of a regression



Hypothesis Test: the overall model

$$F = \frac{MSR}{MSE}$$

Is the model significant?

ANOVA^a

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|-------|------------|----------------|----|-------------|--------|-------------------|
| 1 | Regression | 1839.069 | 1 | 1839.069 | 23.081 | .000 ^b |
| | Residual | 3904.252 | 49 | 79.679 | | |
| | Total | 5743.322 | 50 | | | |

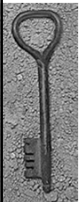
a. Predictors: (Constant), Poverty

b. Dependent Variable: CrimeRate

df for the Model = # of independent variables

df for Residual = n-2, two estimates for the model (a,b)

df for total = n-1



Summary

- ◆ Estimation of linear regression model
- ◆ Concepts: error, covariance, R-square
- ◆ Hypothesis test:
 - Correlation coefficient, constant, coefficient, the overall model
- ◆ SPSS application



Example:

- ◆ Is there a linear association between housing price and number of bedrooms?
- ◆ What kind of linear association is there between housing price and number of bedrooms?