

# Finding Relationships Among Variables

MGMT 662: Integrative Research Project

August 14, 2008.

## 1

### Goals of this section

- Learn how detect linear relationships between variables.
- Learn how to detect relationships between ordinal and categorical variables.
- Learn how to estimation the relationship between many variables.

## 2 Chi-Squared Test for Independence

### 2.1 Contingency Table

#### Chi-Squared Test for Independence

- Used to determine if two categorical variables are related.
- Example: mortality rates on the Titanic:

	Men	Women	Boys	Girls	Total
Survived	332	318	29	27	<b>706</b>
Died	1360	104	35	18	<b>1517</b>
Total	<b>1692</b>	<b>422</b>	<b>64</b>	<b>45</b>	<b>2223</b>

- Data in the table are always frequencies that fall into individual categories.
- Could use this table to test if two variables are independent.

## 2.2 Hypothesis Test

### Test of independence

- **Null hypothesis:** there is no association between the row variable and the column variable.
- **Alternative hypothesis:** The two variables are dependent.
- Test statistic:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- $O$ : observed frequency in a cell from the contingency table.
- $E$ : expected frequency assuming the row and column variable are independent.

$$E = \frac{(\text{row total})(\text{column total})}{\text{grand total}}$$

- **Chi-squared distribution:** a distribution skewed to the right whose support is always positive.

## 2.3 Example

### Example: 2000 Florida Presidential Election

- Data on all but two counties in Florida on the voting technology used and whether Bush had more votes than Gore in the given county.
- Open `floridaelection.xls` in SPSS.
- Click on **Analyze** menu, select **Descriptive Statistics** then **Crosstabs...**
- Put **Technology** in either the row variable or column variable, and put **Bush Winner** in the other.
- Click the **Statistics** button, select the **Chi-square** check box, and click **Continue**.
- Click the **Cells** button, select the **Expected** check box, and click **Continue**.
- Click **OK**. The results show the observed and expected contingency tables as well as the results to the Chi-squared test (but you want to use 1-tailed significance).

## 3 Correlation

### 3.1 Correlation coefficient

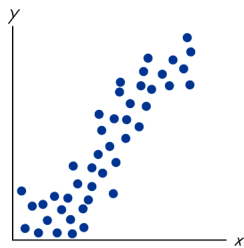
#### Correlation

- A **correlation** exists between two variables when one of them is related to the other in some way.
- The **linear correlation coefficient** is a measure of the strength of the linear relationship between two variables.

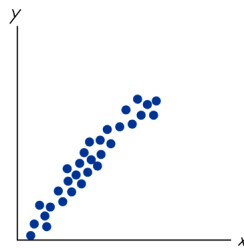
$$\text{Population: } \rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$\text{Sample: } r = \frac{s_{xy}}{s_x s_y}$$

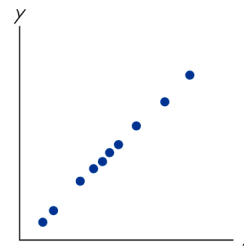
#### Positive linear correlation



(a) Positive correlation between  $x$  and  $y$



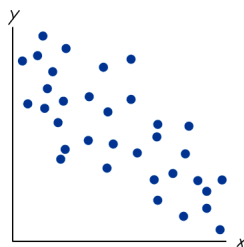
(b) Strong positive correlation between  $x$  and  $y$



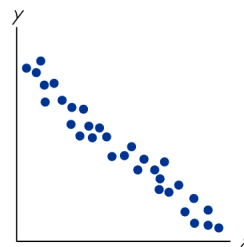
(c) Perfect positive correlation between  $x$  and  $y$

- Positive correlation: two variables move in the same direction.
- Stronger the correlation: closer the correlation coefficient is to 1.
- Perfect positive correlation:  $\rho = 1$

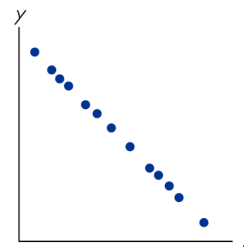
#### Negative linear correlation



(d) Negative correlation between  $x$  and  $y$



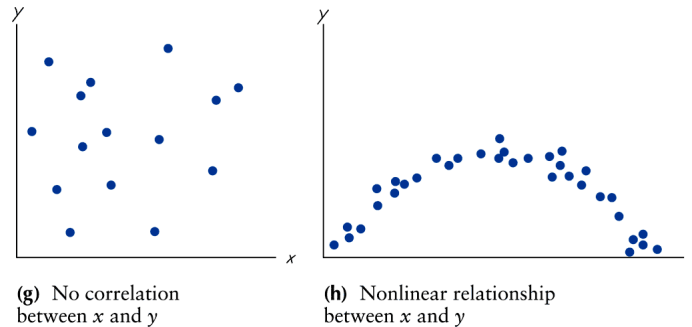
(e) Strong negative correlation between  $x$  and  $y$



(f) Perfect negative correlation between  $x$  and  $y$

- Negative correlation: two variables move in opposite directions.
- Stronger the correlation: closer the correlation coefficient is to -1.
- Perfect negative correlation:  $\rho = -1$

### No linear correlation



- Panel (g): no relationship at all.
- Panel (h): strong relationship, but not a *linear* relationship.
  - Cannot use regular correlation to detect this.

## 3.2 Example

### Example: Public Expenditure

- Data from 1960! about public expenditures per capita, and variables that may influence it:
  - Economic Ability Index
  - Percentage of people living in metropolitan areas.
  - Percentage growth rate of population from 1950-1960.
  - Percentage of population between the ages of 5-19.
  - Percentage of population over the age of 65.
  - Dummy variable: Western state (1) or not (0).
- Is there a statistically significant linear correlation between the percentage of the population who is young and the public expenditure per capita?
- Is there a statistically significant linear correlation between the public expenditure per capita and whether or not the state is a western state?
- Is there another way to test for this relationship that you have already learned? (Hint: yes)

## Using SPSS

- Open `publicexp.xls` in SPSS.
- Go to **Analyze** menu, select **Correlate**, select **Bivariate**.
- Throw the two variables you want to estimate the correlation with into the **Variables** text box.
- Make sure **Pearson Correlation Coefficient** is selected.
- Click on **Options** if you want to also compute standard deviations, variances, and covariances.
- Click **OK!**
- Presented with a redundant table of correlation coefficients and p-values.

## 4 Regression

### 4.1 Regression line

#### Regression

- Regression line: equation of the line that describes the linear relationship between variable  $x$  and variable  $y$ .
- Need to assume that one variable causes another.
  - $x$ : *independent* or *explanatory* variable.
  - $y$ : *dependent* variable.
  - Variable  $x$  can influence the value for variable  $y$ , but not vice versa.
- Example: Suppose one wants to estimate how much smoking affects lung capacity.
  - $x_i$ : quantity of cigarettes smoked per day by individual  $i$  (independent).
  - $y_i$ : lung capacity of individual  $i$  (dependent).

#### Regression line

- Population regression line:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- The actual coefficients  $\beta_0$  and  $\beta_1$  describing the relationship between  $x$  and  $y$  are unknown.
- Use sample data to come up with an estimate of the regression line:

$$y_i = b_0 + b_1 x_i + e_i$$

- Since  $x$  and  $y$  are not perfectly correlated, still need to have an error term.

### Predicted values and residuals

- Given a value for  $x_i$ , can come up with a **predicted value** for  $y_i$ .

$$\hat{y}_i = b_0 + b_1 x_i$$

- This is not likely be the actual value for  $y_i$ .
- **Residual** is the difference between the actual value of  $y_i$  and the predicted value,  $\hat{y}$ .

$$e_i = y_i - \hat{y} = y_i - b_0 - b_1 x_i$$

### Least Squares Estimate

- How should we obtain the “best fitting line”.
- Ordinary least squares (OLS) method.
- Choose sample estimates for the regression coefficients that minimizes:

$$\sum_{i=0}^n (y_i - \hat{y}_i)^2$$

## 4.2 Interpreting the slope

### Interpreting the slope

- Interpreting the slope,  $\beta_1$ : amount the  $y$  is predicted to increase when increasing  $x$  by one unit.
- When  $\beta_1 < 0$  there is a negative linear relationship. That is increasing  $x$  causes  $y$  to decrease.
- When  $\beta_1 > 0$  there is a positive linear relationship. That is increasing  $x$  causes  $y$  to increase.
- When  $\beta_1 = 0$  there is no linear relationship between  $x$  and  $y$ .

### Example: Public Expenditure

- Data from 1960 about public expenditures per capita, and variables that may influence it.
- In SPSS, choose **Analyze** menu and select **Regression** and **Linear**.
- Select **EX** (Expenditure per capita) as your dependent variable. This is the variable your are interested in explaining.
- Select your independent (aka explanatory) variables. These are the variables that you think can explain the dependent variable. I suggest you select these:

- ECAB: Economic Ability
- MET: Metropolitan
- GROW: Growth rate of population
- WEST: Western state = 1.

### Example: Public Expenditure

- If the percentage of the population living in metropolitan areas is expected to increase by 1%, what change should we expect in public expenditure?
- Is this change statistically significantly different from zero?
- Accounting for economic ability, metropolitan population, and population growth, how much more do Western states spend on public expenditure per capita?

## 4.3 Variance Decomposition

### Sum of Squares Measures of Variation

- **Sum of Squares Regression (SSR)**: measure of the amount of variability in the dependent (Y) variable that is explained by the independent variables (X's).

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

- **Sum of Squares Error (SSE)**: measure of the unexplained variability in the dependent variable.

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- **Sum of Squares Total (SST)**: measure of the total variability in the dependent variable. Does the formula below look familiar?

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

- SST = SSR + SSE.

### Coefficient of determination

- The **coefficient of determination** is the percentage of variability in  $y$  that is explained by  $x$ .

$$R^2 = \frac{SSR}{SST}$$

- $R^2$  will always be between 0 and 1. The closer  $R^2$  is to 1, the better  $x$  is able to explain  $y$ .
- The more variables you add to the regression, the higher  $R^2$  will be.
- The Adjusted  $R^2$  penalizes additional variables.

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-k-1} (1 - R^2)$$

- When the adjusted  $R^2$  increases when adding a variable, then the additional variable really did help explain the dependent variable.
- When the adjusted  $R^2$  decreases when adding a variable, then the additional variable does not help explain the dependent variable.

### F-test for Regression Fit

- F-test for Regression Fit: Tests if the regression line explains the data.
- Very, very, very similar to ANOVA F-test.
- $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$ .
- $H_1$  : At least one of the variables has explanatory power (i.e. at least one coefficient is not equal to zero).

$$F = \frac{SSR/(k-1)}{SSE/(n-k)}$$

- Where  $k$  is the number of explanatory variables.

### Example: Public Expenditure

- In the previous example, how much of the variability in public expenditure is explained by the following four variables:
  - ECAB: Economic Ability
  - MET: Metropolitan
  - GROW: Growth rate of population
  - WEST: Western state = 1.

- Is the combination of these variables significant in explaining public expenditure?
- Re-run the regression, this time also including:
  - YOUNG: Percentage of population that is young.
  - OLD: Percentage of population that is old.

**Example: Public Expenditure**

- What happened to the coefficient of determination?
- What happened to the adjusted coefficient of determination? What is your interpretation?
- What happened to the estimated effect of the other variables: metropolitan area? Western state?