社会统计学及SPSS软件应用

Instructor:王荣欣

Email: rxwang@qq.com

周一3-4节、单周周四3-4节，3A106-2

2020年10月12日

CONTENTS

1 回归分析

- 回归分析概述
- 回归分析的基本思想
- 一元线性回归

回归分析

1. 普通线性回归：因变量Y必须是连续型数据
2. 0-1回归：因变量Y是0-1型数据
3. 定序回归：因变量Y是定序数据
4. 计数回归：因变量Y是数数的数据（非负的整数）
5. 生存数据回归：因变量Y是生存数据

WHAT IS Y AND X?

Y，称为因变量，即因为别人的改变而改变的变量。

X就是用来解释Y的相关变量，可以是一个，也可以是很多个。

GALTON AND REGRESSION

- Regression paradox: tall parents may expect to have less tall children, and tall children tend to have less tall parents.

- Francis Galton(1822-1911): *Regression towards mediocrity in hereditary stature* (1886)

*教材第84页

There are three types of variables in a regression model: a dependent variable, a set of independent variables, and random errors.

1. A dependent variable represents a population characteristic of interest being explained in a study.
2. A independent variable are used to explain the variation in the dependent variable.
3. The dependent variable depends on, or is explained by independent variables in a regression-type statistical models.

LINEAR REGRESSION

1. For any two variables, say X and Y, if we know the value of X, what does this tell us about Y?
2. What can be said about the conditional distribution of Y, given X. Given X, are we better able to predict the mean or median of Y?
3. Does the variation in Y change with X? How might we characterize the strength of any association that might exist?

回归分析就是追本溯源

- 就是把因变量的“变化”回溯或归根到自变量的变化，
- 或者说，就是探索因变量的变化是否（以及如何）与其他变量的变化相关（correlated）。

- 回归分析的任务就是通过研究X和Y的相关关系，尝试去解释Y的形成机制，进而达到通过X去预测Y的目的。
- 回归分析的目的是利用变量间的简单函数关系，用自变量对因变量进行“预测”，让“预测值”尽可能地接近因变量的“观测值”。

回归分析的特点就在于它把观测值分解成两部分：结构部分和扰动部分。

观测项=结构项+随机项
observed part=structural part + stochastic part

1. 观测项部分代表因变量的实际取值；
2. 结构项部分代表因变量和自变量之间的结构关系，表现为“预测值”；
   - The structural part denotes the relationship between the dependent and independent variable.

随机项部分表示观测项中未被结构项解释的剩余部分。

- The stochastic part is the random component unexplained by the structural part.

随机项包含三部分：

1. omitted structural factors 被忽略的结构因素（包括结构项的差错）
2. measurement error 测量误差（由数据测量、记录或报告过程中的不精确导致）
3. random noise 随机干扰（反映了人类行为或社会过程不可避免地受到不确定性因素的影响）

观测项=Structural part+ stochastic part

观测项=0%+100% （纯粹扰动）

观测项=100%+0% （完全决定）

观测项=概括项+残差项

Observed=summary+residual

- The model serves to summarize the basic feature of data. The question is whether the model corresponds to the facts.
- 统计模型的主要目标在于用最简单的结构和尽可能少的参数来概括大量数据所包含的主要信息。
    - The primary goal of statistical modeling is to summarize massive amounts of data with simple structures and few parameters.

REGRESSION EQUATION

- In 1925, Fisher introduced $Y = a + bX$ and incorporated the terms ”dependent” variable and ”independent” variable.
- Fisher produced the equation for the regression (or predicted) line: $\hat{Y} = a + bX$.

- Simple Linear regression 一元线性回归模型可以表示为：

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

1. Y_i表示第i名个体在因变量Y上的取值
2. X_i表示第i名个体在自变量X上的取值
3. β_0、β_1是模型的参数，需要根据样本数据进行估计
4. ϵ是随机误差项

- Multiple Linear regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \epsilon_i$$

SIMPLE LINEAR REGRESSION

1

1 Imagine we have n pairs of values

$$(X_1, Y_1), ...(X_n, Y_n)$$

2 The immediate goal is to find a straight line for predicting Y given some value for X. The goal is to find a prediction rule having the form

$$\hat{Y} = b_0 + b_1 X$$

3 For each of the observed X values, namely, $X_1, ..., X_n$, we have a predicted Y value:

$$\hat{Y}_i = b_0 + b_1 X_i$$

There will be some discrepancy between the observed Y_i and its predicted value, $\hat{Y}_i$. This discrepancy van be measured with

$$r_i = Y_i - \hat{Y}_i$$

which is called a residual. Residuals simply represent the error in our prediction rule based on $\hat{Y}$.

The least squares principle for determining the slope and intercept is to determine the values for b_0 and b_1 that minimize the sum of the squared residual.

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2 \text{（最小化）}$$

$$\frac{\partial D}{\partial b_0} = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0 \text{ (1)}$$

$$\frac{\partial D}{\partial b_1} = -2\sum_{i=1}^{n}X_i(Y_i - b_0 - b_1 X_i) = 0 \text{ (2)}$$

$$nb_0 + b_1\sum_{i=1}^{n}X_i = \sum_{i=1}^{n}Y_i \text{ (1)}$$

$$b_0\sum_{i=1}^{n}X_i + b_1\sum_{i=1}^{n}X_i^2 = \sum_{i=1}^{n}(X_i Y_i) \text{ (2)}$$

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$b_0 = \bar{Y} - b_1\bar{X}$$

Suppose we have a set of data arrayed like this:

Father’s Years of Schooling	Respondent’s Years of Schooling
2	4
12	10
4	8
13	13
6	9
6	4
8	13
4	6
8	6
10	11

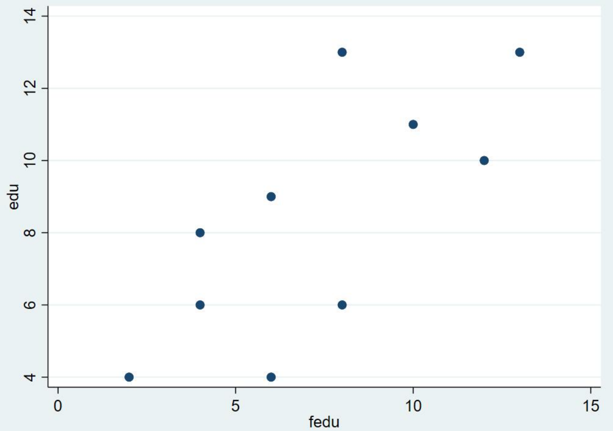
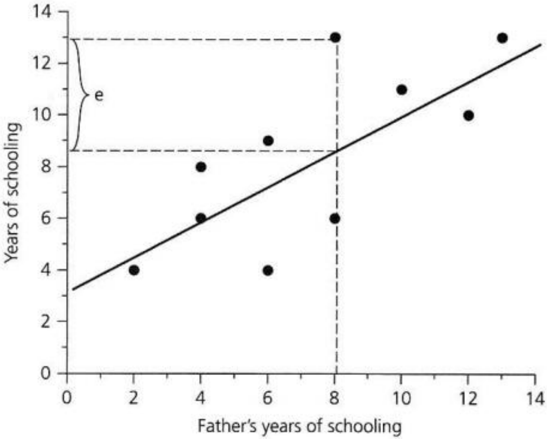


Figure 4.1: 散点图

Example 4.1

利用该数据，研究教育的代际流动：

1. 判断最佳的拟合直线方案
2. 计算直线的拟合优度
3. 检验数据是否支持受教育年限受到父亲受教育年限的影响这一研究假设（显著度$\alpha = 0.05$）
4. 在*95%*置信水平下，预测父亲受教育年限为*10*年者的平均受教育年限。


14
12
10
8
6
4
2
0
e
Years of schooling
0
2
4
6
8
10
12
14
Father's years of schooling

回归模型的系数

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

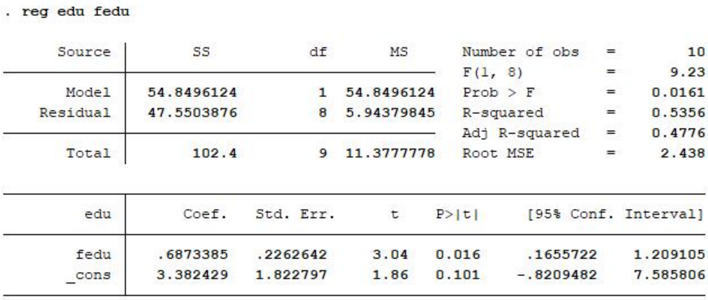
或$b_1 = \frac{Cov(x,y)}{(S_x)^2} = \frac{covariance\ \ of\ \ x\ \ and\ \ y}{variance\ \ of\ \ x}$

1. $\bar{X} = 7.3, \bar{Y} = 8.4$
2. $\sum_{i=1}^{n}(x_i - \bar{X})^2 = 116.1, \sum_{i=1}^{n}(y_i - \bar{Y})^2 = 102.4$
3. $\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}) = 79.8$
4. $b_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})}{\sum_{i=1}^{n}(x_i-\bar{X})^2} = 79.8/116.1 = 0.687$
5. $b_0 = \bar{Y} - b_1\bar{X} = 8.4 - 0.687 * 7.3 = 3.38$

$$R^2 = \frac{[\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})]^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

$$= \frac{79.8^2}{116.1 \times 102.4}$$

$$= 0.536$$



```
. reg edu fedu

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(1, 8)       =    9.23
       Model |  54.8496124     1  54.8496124           Prob > F      =  0.0161
    Residual |  47.5503876     8  5.94379845           R-squared     =  0.5356
-------------+------------------------------           Adj R-squared =  0.4776
       Total |       102.4     9  11.3777778           Root MSE      =   2.438

------------------------------------------------------------------------------
         edu |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        fedu |   .6873385   .2262642     3.04   0.016     .1655722    1.209105
       _cons |   3.382429   1.822797     1.86   0.101    -.8209482    7.585806
------------------------------------------------------------------------------
```

Figure 4.2: 父亲受教育年限与受访者受教育年限的关系

1 最佳的拟合直线Y=3.382+0.687X。
2 $R^2 = 0.536$说明：回归方程能够解释平均受教育年限总方差中的53.6%。
3 支持受教育年限受到父亲受教育年限的影响这一研究假设。
4 当父亲受教育年限为10年, 预测的平均受教育年限为：
$E(edu|fedu = 10) = \hat{\beta}_0 + \hat{\beta}_1 x = 3.38 + 0.687 * 10 = 10.25$

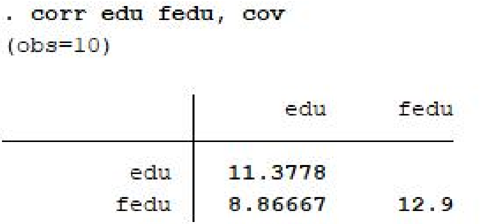
COVARIANCE（协方差）

- 协方差用于测量两个变量之间的线性关系。
- The covariance is the measure of how much two random variables move together.

$$Cov(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})}{n-1}$$

1. If two variables tend to move together in the same direction, then the covariance between the two variables will be positive.
2. If two variables move in the opposite direction, the covariance will be negative.
3. If there is no tendency for two variables to move one way or the other, then the covariance will be zero. （X和Y彼此独立，有Cov(X,Y)=0）

父亲受教育年限与受访者受教育年限的协方差：（方差是协方差的特例，X的方差就是X与其自身的协方差）



```
. corr edu fedu, cov
(obs=10)

             |      edu     fedu
-------------+------------------
         edu |  11.3778
        fedu |  8.86667     12.9
```

regression slope $b_1 = \frac{Cov(x,y)}{(S_x)^2} = \frac{covariance\ of\ x\ and\ y}{variance\ of\ x} = \frac{8.87}{12.9} = 0.687$

Pearson’s Product-Moment Correlation，即correlation coefficient（相关系数）

- Pearson introduced the term simple correlation when measuring a linear relationship between two continuous variables only, such as the relationship between stature of father and stature of son.

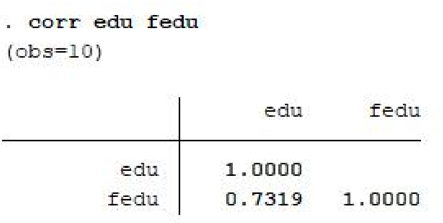
$$r = \frac{Cov(x,y)}{(S_x)(S_y)} = \frac{covariance}{(standard\ \ deviation\ \ of\ \ x)(standard\ \ deviation\ \ of\ \ y)}$$

1 相关有正有负：correlation coefficient取值范围：-1到1。

2 相关有强有弱：大于0.7为强相关。

3 相关有显著有不显著（significant）。

   - 显著度检验：此变量的变化与彼变量的变化有systematic, non-random的关系。



```
. corr edu fedu
(obs=10)

             |      edu     fedu
-------------+------------------
         edu |   1.0000
        fedu |   0.7319   1.0000
```

1 当X和Y彼此独立，有Cov(X,Y)=0，r=0。
2 r=0表示X和Y之间不存在线性关系。
3 协方差是有量纲的，相关系数没有量纲，相关系数可以进行直接比较。

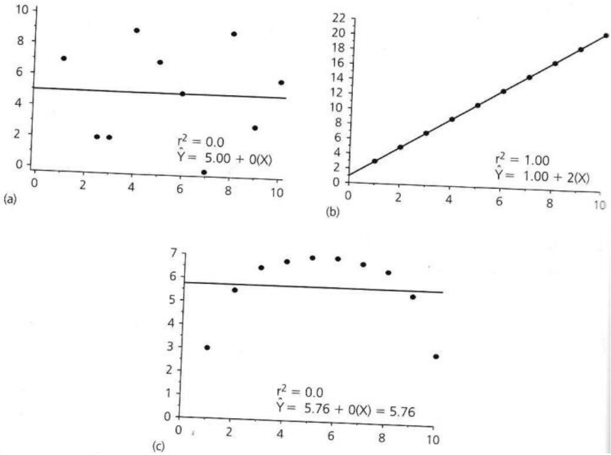


Figure 4.3: 教材p.88

1. The square of the Pearson correlation coefficient $r^2 = 0.7319^2$= 0.536, which tells us that the variance around the regression line is about half the size of variance in educational attainment is explained by the corresponding variability in father’s education.
2. 判定系数（拟合优度）R^2实际上是Pearson相关系数的平方。

参考文献

- 谢宇，2013，《回归分析》（修订版），北京：社会科学文献出版社。