STAT 417 Lecture Note 24

In previous lectures, we used χ^2 - test to study the relationships among discrete variables. The χ^2 - test is only useful for discrete data. When data is continuous. χ^2 - test is no longer useful. A more powerful approach is needed.

Linear Regression:

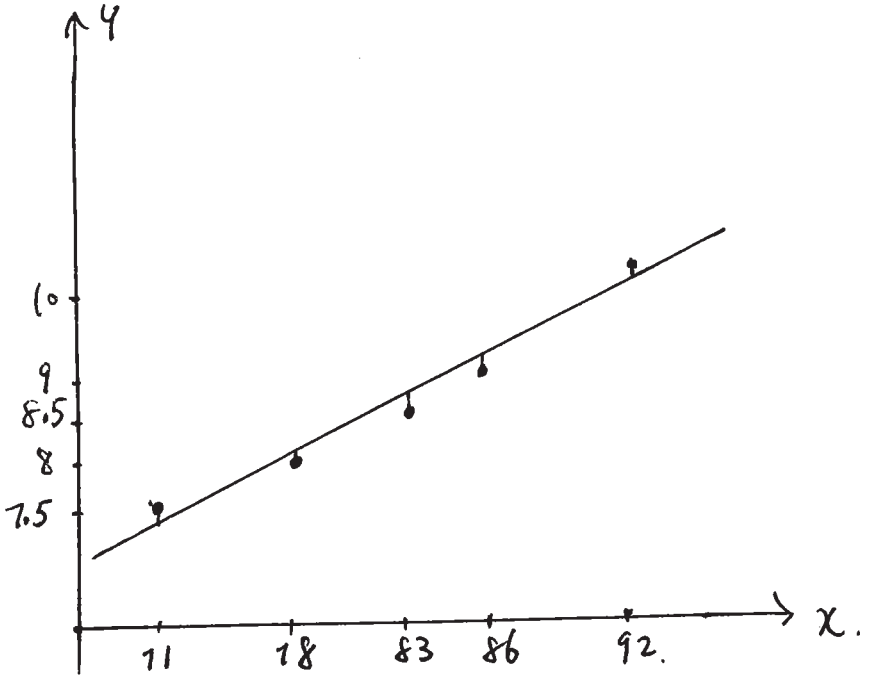
$n = 5$ students graduated last summer. below are their grades in stat 417 (X) and current salaries (Y) in \$10,000 :

X	92	86	83	78	71
Y	10	9	8.5	8	7.5

Question: does STAT417 grades affect your future salary?

Both grades and salaries are continuous. so χ^2 - test is no longer useful. We will use linear regression technique to study the relationship between X and Y.

understanding the data:



From the above plot, it is easy to see that Y is increasing with X. we want to establish a linear relationship between x and y. the idea is to find a line $y = b_1 + b_2 x$ to fit the data as good as possible.

$$SE = \sum_{i=1}^{5} (Y_i - b_1 - b_2 x_i)^2$$

want to find b_1, b_2 to minimize SE.

$$\frac{\partial}{\partial b_1} SE = -\sum_{i=1}^{5} 2(Y_i - b_1 - b_2 x_i) = 0 \Rightarrow b_1 = \frac{1}{5}\sum_{i=1}^{5} (Y_i - b_2 x_i)$$

$$= \bar{Y} - b_2 \bar{X}$$

$$\frac{\partial}{\partial b_2} SE = -\sum_{i=1}^{5} 2x_i(Y_i - b_1 - b_2 x_i) = 0.$$

$$\Rightarrow -\sum_{i=1}^{5} 2x_i(Y_i - \bar{Y} + b_2\bar{x} - b_2 x_i)$$

$$= -\sum_{i=1}^{5} 2x_i(Y_i - \bar{Y} - b_2(x_i - \bar{x})) = 0$$

$$\text{so} \quad b_2 = \frac{\sum_{i=1}^{5} x_i(Y_i - \bar{Y})}{\sum_{i=1}^{5} x_i(x_i - \bar{x})}$$

Ex1: show: $\sum_{i=1}^{5} x_i(Y_i - \bar{Y}) = \sum_{i=1}^{5} (x_i - \bar{x})(Y_i - \bar{Y})$.

$$\sum_{i=1}^{5} x_i(x_i - \bar{x}) = \sum_{i=1}^{5} (x_i - \bar{x})^2.$$

proof: $$\sum_{i=1}^{5} (x_i - \bar{x})(Y_i - \bar{Y}) = \sum_{i=1}^{5} x_i(Y_i - \bar{Y}) - \sum_{i=1}^{5} \bar{x}(Y_i - \bar{Y})$$

$$= \sum_{i=1}^{5} x_i(Y_i - \bar{Y}).$$

Note: $\sum_{i=1}^{5} \bar{x}(Y_i - \bar{Y}) = \bar{x}\sum_{i=1}^{5}(Y_i - \bar{Y}) = \bar{x}\left(\sum_{i=1}^{5} Y_i - 5\bar{Y}\right) = 0.$

$$\sum_{i=1}^{5} \bar{x}(x_i - \bar{x}) = \bar{x}\sum_{i=1}^{5}(x_i - \bar{x}) = \bar{x}\left(\sum_{i=1}^{5} x_i - 5\bar{x}\right) = 0$$

so $$\sum_{i=1}^{5} (x_i - \bar{x})^2 = \sum_{i=1}^{5} x_i(x_i - \bar{x}) - \sum_{i=1}^{5} \bar{x}(x_i - \bar{x}) = \sum_{i=1}^{5} x_i(x_i - \bar{x}).$$

By Ex1. we get the following

$$b_{2:} = \frac{\sum_{i=1}^{5} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{5} (x_i - \bar{x})^2} = 0.118.$$

So. $b_1 = \bar{y} - b_{2:}\bar{x} = -1.076.$

So. the line we want is $y = -1.076 + 0.118x.$

this line is called the "line of best fit".

General case: Suppose we have n samples

$$(x_1, y_1),\ (x_2, y_2),\ \cdots,\ (x_n, y_n).$$

The line of best fit, or formally, the best fitting line, is

$$y = b_1 + b_2 x,$$

where
$$\begin{cases} b_2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ b_1 = \bar{y} - \beta_2 \bar{x} \end{cases}$$

b_1, b_2 are the estimators of the intercept β_1 and slope β_2.

prediction of future ~~[illegible]~~ salary.

Suppose the 6th graduate student has STAT 417 grade 88. then how much s/he will earn?

Recall the ~~fitted~~ line of the best fit is

$$y = -1.076 + 0.118 x.$$

when $x = 88$.

$$y = 9.3$$

So ~~the one~~ it is predicted that the future salary is ~~[illegible]~~ $93,000

Theorem 1.

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = b_2^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} (Y_i - b_1 - b_2 X_i)^2 .$$

proof:

$$\sum_{i=1}^{n} (Y_i - b_1 - b_2 X_i)^2 = \sum_{i=1}^{n} \left(Y_i - \bar{Y} - b_2 (X_i - \bar{X}) \right)^2$$

$$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - 2 \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}) b_2 + b_2^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - 2 \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2} b_2 \cdot \sum (X_i - \bar{X})^2$$

$$+ b_2^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - 2 b_2^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 + b_2^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - b_2^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 .$$

If Y and X have a strong relationship, then b_2 should be far away from zero (i.e. b_2^2 is large). this means $b_2^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$ should take a large proportion of $\sum (Y_i - \bar{Y})^2$.

Does your future salary depend on your STAT 417 score?

An ANOVA test. ANOVA = analysis of variance.

use the following so called ANOVA Table:

Source	df	Sum of squares	Mean squares
X	1	$b_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$	$b_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$
Error	n-2	$\sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2$	$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2$
Total	n-1	$\sum_{i=1}^{n} (y_i - \bar{y})^2$	

F-test: F is named after Ron Fisher.

$$F = \frac{b_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{s^2} \approx F(1, n-2).$$

Intuitively, a larger F value means a stronger relationship between x and y.

The distribution of F can be used to determine if F is big or small.

In our grade - salary example, the ANOVA table is

Source	df	Sum of Squares (SS)	MS
X	1	3.54	3.54
Error	3	0.16	0.053
total	4	3.7	

so $F = \frac{3.54}{0.053}$

$= 65.6$

$> F_{0.95}(1, 3)$

$= 10.13.$

STAT 417 Final review:

A brief review:

1. ~~posterior~~ Bayes model: { ~~Optim~~ Bayesian estimation { posterior expectation; posterior mode.
   ~~Bayes~~ credible interval. (only for normal model)
   Bayesian hypothesis testing { prior odds ratio; posterior odds ratio; Bayes factor.

2. model check: by χ^2-test. { check normal model. check. multinomial model. check. exponential model

3. independence check. { Check independence of among. Categorical variables. by χ^2-test
   check independence among continuous variables by linear regression

4. Linear regression { ~~best fit~~ line. $y = b_1 + b_2 x$. { $b_1 = \cdots$; $b_2 = \cdots$
   line of the best fit.
   ~~ANOVA test~~ prediction: $y_{new} = b_1 + b_2 x_{new}$.
   ANOVA Test of if Y and X are independent.
   i.e. test H_0: slop = 0