

"Regression" Analysis I ... Descriptive Statistics for SLR

This concerns quantifying relationships between variables ... predicting some response y on the basis of one or more "predictors" (or explanatory variables) x_1, x_2, \dots, x_k

Example What we will fondly refer to as "the Real Estate Example" concerns a data set from a 1986 *Real Estate Appraiser* article. We will initially concern ourselves with

y = selling price of a home (\$1000)
 x = home size (100 ft²)

There is also a "home condition" variable we'll use later.

1

Selling Price	Size	Condition
60.0	23	5
32.7	11	2
57.7	20	9
45.5	17	3
47.0	15	8
53.3	21	4
64.5	24	7
42.6	13	6
54.5	19	7
57.5	25	2

The simplest possible version of this "regression" problem is where

1. there is a single x to be used in predicting/explaining y

2

2. the relationship between y and x is "approximately straight line"

This is "Simple Linear Regression" (SLR). The text treats this in Ch 2 (on a descriptive level) and in Ch 10 (on an inferential level).

Ch 2 deals primarily with 2 questions:

1. What is "the best" line to use in describing how y varies with x ?
2. How may one quantify the strength of any apparent linear relationship between x and y ?

3

Fitting a Line by "Least Squares"

The traditional (mathematically simple) answer to "How do I choose a 'best fitting' line to describe an (x, y) data set?" is to apply the principle of "least squares." This involves looking for a line that makes as small as possible the sum of squared vertical distances from plotted (x, y) points to the line. Pictorially, this means that one jiggles the line on the plot below until one has made the sum of shaded areas as small as possible.

4

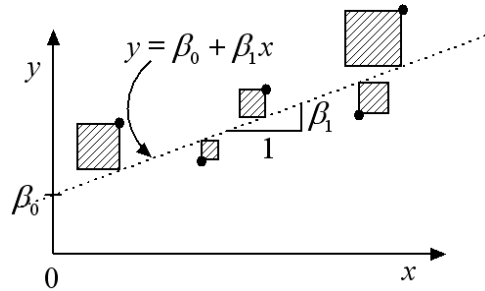


Figure 1: Fitting a Line by "Least (Sum of) Squares"

5

Algebraically, the problem of fitting a line is to choose a slope β_1 and an intercept β_0 to make as small as possible the quantity

$$\sum (y_i - (\beta_0 + \beta_1 x_i))^2$$

This is a calculus problem that has a nice simple solution. As it turns out, the "best" slope is

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

and then the "best" intercept is

$$b_0 = \bar{y} - b_1 \bar{x}$$

These are called the least squares slope and intercept.

Example In the Real Estate Example, one can use the Fit Y by X option in JMP and find that

$$b_1 = 1.9001 \text{ (\$1000/100 ft}^2\text{)} \text{ and } b_0 = 16.008 \text{ (\$1000)}$$

6

that is, the least squares line is

$$y = 16.008 + 1.9001x$$

for y (price) in \$1000 and x (size) in 100 ft². We judge that in 1986, price of homes in the area under consideration increased at a rate of about \$19/ft². The JMP report is

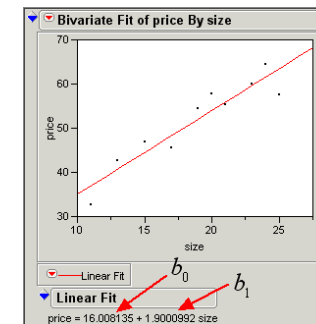


Figure 2: JMP Report for the Real Estate Data

7

8

Exercise For the small fake data set (below and on the handout) find the best fitting line "by hand" (using at most a pocket calculator).

x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
-2	4					
-1	3					
0	3					
1	1					
2	-1					

Measuring Strength of Linear Association

As an initial measure of association between two variables x and y we'll discuss the so-called "sample correlation" ("sample *linear* correlation")

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

MMD&S present an alternative (equivalent) formula for this, namely

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Example As it turns out, in the real estate example

$$r = .9379$$

(This can be obtained easily from JMP using Analyze→Multivariate Methods →Multivariate or otherwise.) The obvious question is then, what does .94 mean?

Some basic facts about the sample correlation are:

1. $-1 \leq r \leq 1$ always!
2. $r = 1$ exactly when all (x, y) points fall on a single straight line with positive slope, and $r = -1$ exactly when all (x, y) points fall on a single straight line with negative slope ... in this regard, compare the formulas for b_1 and r (they are very much alike) and note that they must have the same sign

People then interpret r "near" ± 1 as indicating "strong linear relationship" between x and y , and r "near" 0 as indicating "lack of linear relationship."

Example In the Real Estate Example, $r = .94$ indicates a fairly strong relationship between *size* and *price*.

Exercise Compute r for the small fake data set "by hand" (using at most a pocket calculator).

x	y	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
-2	4					
-1	3					
0	3					
1	1					
2	-1					

A second measure of strength of linear association between x and y (that turns out to be related to the first) is the "coefficient of determination" or R^2 . This is built on two measures of variability in the y 's called "sums of squares." The first is a "total" sum of squares

$$SSTot = \sum (y_i - \bar{y})^2$$

$$= (n - 1) s_y^2$$

(Remember that in Session 1 we remarked that simple algebra starting with the definition of the sample variance makes $(n - 1) s^2 = \sum (x_i - \bar{x})^2$.) This is a measure of "raw variability in y " since it is "most of" the sample variance of the y 's.

A second sum of squares is a measure of variation "not accounted for" or "left over" or "residual" after fitting an equation to a data set. Let

\hat{y}_i = the prediction of response for case i from a fitted equation

13

(in the present context this is the y -value on the fitted line for the i th value of x). Then define

$$e_i = y_i - \hat{y}_i$$

= a measure of what's unaccounted for by the fitted equation in case i

= a measure of what's "left over"/residual in case i

The sum of squares of these is the "error" or "residual" sum of squares

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

This is a measure of variability in response not accounted for by the fitted equation. For the equations we'll fit in Stat 328

$$SSE \leq SSTot$$

For the small fake data set, this can be illustrated graphically as

14

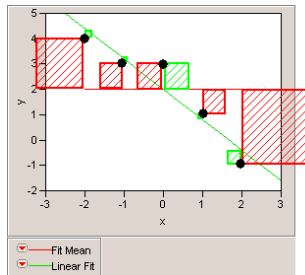


Figure 3: $SSTot$ (in red) and SSE (in green) for the Fake Data Set

15

Since $SSE \leq SSTot$, the difference

$$SSR = SSTot - SSE$$

is a measure of variation in y that is accounted for by the fitted equation. It is called a "regression sum of squares." It is common to state this as a fraction of $SSTot$ and to define

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

That is, R^2 is a fraction of the overall/raw variability in y in some sense accounted for in fitting the equation. Always

$$0 \leq R^2 \leq 1$$

Example In the Real Estate Example, the JMP Fit Y by X routine prints out R^2 and here

$$R^2 = .8796$$

16

88% of the raw variability in price can be accounted for or "explained" using home size as a predictor/explanatory variable.

Exercise Compute R^2 for the small fake data set "by hand" (using at most a pocket calculator).

x	y	$\hat{y} = 2 - 1.2x$	$e = y - \hat{y}$	$e^2 = (y - \hat{y})^2$
-2	4			
-1	3			
0	3			
1	1			
2	-1			

As it turns out, R^2 can be thought of as a squared sample correlation. That is, it is ALWAYS the case that

$$\frac{SSR}{SStot} = R^2 = (\text{correlation between } y \text{ and } \hat{y})^2$$

IN THE SPECIAL CASE of SLR we are now studying, \hat{y} and x are perfectly correlated and

$$(\text{correlation between } y \text{ and } \hat{y})^2 = (\text{correlation between } y \text{ and } x)^2$$

so that in this case $R^2 = r^2$.

Example In the Real Estate Example, notice that

$$R^2 = .8796 = (.9379)^2 = r^2$$

Notice then that in a SLR problem, if one is furnished with a value for R^2 , one can infer the sample correlation between x and y to be

$$r = \pm\sqrt{R^2}$$

where the correct sign is the sign of the slope of the least squares line, b_1 .

Several important caveats about the practical use of SLR (and more complicated forms of regression, for that matter) are:

1. the least squares line and r (R^2) are *highly* sensitive to a few "extreme" data points
2. r and R^2 measure only *linear* association
3. one should/must beware extrapolation
4. correlation is not necessarily causation

Regarding caveat number 1) below is a real scatterplot of ages and heights in class of Purdue undergraduates. With the tall older student in the data set, $r = .73$. If that person is removed, $r = .03$.

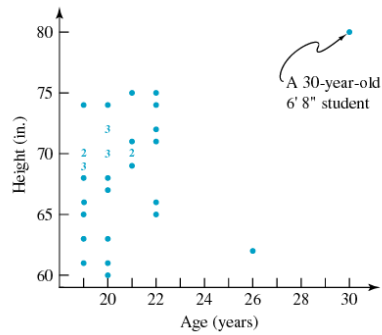


Figure 4: Ages and Heights of 36 Purdue Undergraduates

Regarding caveat number 2) consider the small fake data set

x	y
-2	4
-1	1
0	0
1	1
2	4

There is a *perfect* (non-linear) *relationship* between x and y , namely $y = x^2$. However, you may verify that for this fake data set $r = 0$.

Regarding caveat 3), there is no guarantee that a relationship between x and y established on the basis of data with $13 \leq x \leq 27$ continues to hold for $x = 75$.

Regarding caveat 4), consider a sports example. Points scored and fouls committed by college basketball players are probably highly positively correlated. But would you coach a player to foul a lot in order to improve point production?